

John Benjamins Publishing Company



This is a contribution from *New Perspectives on the Origins of Language*.

Edited by Claire Lefebvre, Bernard Comrie and Henri Cohen.

© 2013. John Benjamins Publishing Company

This electronic file may not be altered in any way.

The author(s) of this article is/are permitted to use this PDF file to generate printed copies to be used by way of offprints, for their personal use only.

Permission is granted by the publishers to post this file on a closed server which is accessible to members (students and staff) only of the author's/s' institute, it is not permitted to post this PDF on the open internet.

For any other use of this material prior written permission should be obtained from the publishers or through the Copyright Clearance Center (for USA: www.copyright.com).

Please contact rights@benjamins.nl or consult our website: www.benjamins.com

Tables of Contents, abstracts and guidelines are available at www.benjamins.com

Brave new words

Pierre J. Bancel & Alain Matthey de l'Étang

Association d'Études linguistiques et anthropologiques préhistoriques, Paris*

Contrary to the received idea that globally spread *papa/mama* words are constantly reinvented by children in different languages, we show here that these words are always inherited from the most ancient stages of their respective families, with the exception of a number of borrowings – which are not innovations, either. We then show that probabilistic calculations aiming to demonstrate that global and other remote etymologies might be mere chance resemblances are invalid, and that chance cannot be reasonably invoked in the cases these calculations deal with. Consequently, the global convergence of *papa/mama* words can only be a trace of a common heritage of all human languages. Finally, we link this finding with others, indicating that these words must have appeared early, most probably at the very origin of articulate language.

1. The Proto-Sapiens kinship terms *papa*, *mama* and *kaka*

Our central claim is that most modern *papa/mama* words, so widespread in all language families worldwide, may be traced back to a common origin. We use the name Proto-Sapiens for the original ancestor language from which they have been inherited, and which must have been the ancestral language of all known languages spoken by modern human beings, who together constitute the species *Homo sapiens*.

On archeological and genetic grounds, Proto-Sapiens may be dated between 200,000 years ago (the approximate earliest date at which our species emerged in Africa; McDougall et al. 2005; White et al. 2003) and 50,000 years ago (the approximate latest date at which our direct ancestors may have left the African continent and began to spread their bones, genes, artifacts, and language over the rest of the world). However, recent archeological findings from several South African sites – Klasies River Mouth (Deacon 2001; Singer & Wymer 1982), Blombos Cave (d'Errico et al. 2005; Henshilwood & d'Errico 2005; Henshilwood et al. 2001) and Pinnacle Point

* Mail should be sent to first author: pierrejbancel@hotmail.com. Thanks to Peter MacNeilage, Claire Lefebvre, and John Bengtson for their useful remarks and corrections, and to Shahar Fineberg and Zofia Laubitz for their fine proofreading job. Errors, of course, are solely ours.

(Marean et al. 2007) – have revolutionized the dating of modern Sapiens anatomy and behavior. Until recently, modern behavior was widely believed to have appeared no earlier than some 40,000 years before present (yBP). But all these sites have revealed numerous unambiguous traces of modern behavior (use of marine food, cooking food on hearths, microlithic tools, polished bone tools, personal ornaments, geometrical engravings, etc.) older than 80,000 yBP, and up to 164,000 yBP at Pinnacle Point. Genetics-based datings like those of the “mitochondrial Eve” around 200,000 yBP (Cann et al. 1987), or the split of Khoisan people between 70,000 to 90,000 yBP (Knight et al. 2003), as well as the antiquity of the first human occupation of Australia and New Guinea (at least 46,000 yBP, and perhaps 60,000 yBP for Australia; see Bowler et al. 2003; O'Connell & Allen 2004) also tend to push Proto-Sapiens back to a date earlier than 50,000 yBP, perhaps as far back as some 100,000 yBP.

1.1 Historical background

The global distribution of *papa/mama* words, noted as early as the mid-nineteenth century (Buschmann 1852), received its currently accepted explanation in the late 1950s. Murdock (1959) and Jakobson (1960) – probably drawing on Lubbock (1889) or Westermarck (1891), though they do not quote any predecessor in this regard – explained that modern *papa/mama* words must be recent and had resulted from constrained, convergent innovations due to child/parent interaction in unrelated languages. In particular, Jakobson claimed that *mama* words derived from the nasal murmur *mmm... mmm...* of suckling babies; he left *papa* words unexplained – but may have considered, from a far-fetched structuralist perspective, that the non-nasal counterpart *p* of nasal consonant *m* should naturally apply to the non-breastfeeding counterpart of the mother, namely the father.

This theory was not supported by any historical evidence. Its authors relied on the growing body of observations of child language acquisition to build an indirect explanation, along the lines of “kinship appellatives resemble each other much too much to have arisen by chance. Since conventional wisdom has it that the many language families they appear in are unrelated to each other, here is how they might have been spontaneously invented in various languages, even though this process has never been observed.” In spite of its indirectness and the good bit of wishful thinking it relied on, this theory immediately received wide approbation and is still taught in linguistics departments as the obvious explanation of the global distribution of *papa/mama* words.

Murdock and Jakobson's view was first challenged 35 years later by the American linguist Merritt Ruhlen (1994a). He had discovered a new widespread appellative *kaka* ‘brother, uncle’, which had escaped the attention of comparatists for a century and a half after the global distribution of *papa/mama* words had become known to linguists. Its phonetic form was unlikely to have emerged from the babbling of babies, since velars like *k* are acquired later than labials (*p, b, m*) and dentals (*t, d, n*). And it seemed

unlikely that its meaning had emerged independently with the same phonetic form in many different language families. Ruhlen concluded that the many *kaka* words he had discovered in a range of language families from Eurasia, the Americas and Oceania had to have been inherited from a common ancestral Proto-Sapiens language. He also suggested that there had to be an inherited component behind the global distribution of *papa/mama* words as well, and that Jakobson's explanation of their origin by convergence was probably "exaggerated, if not completely mistaken" (p. 124).

Ruhlen's discovery prompted us to undertake a global etymological comparison of kinship appellatives. We first checked the etymological support of Proto-Sapiens *kaka*, and found literally hundreds of *kaka* words in parts of the world where they had not been documented by Ruhlen, notably Africa (in the Niger-Congo, Nilo-Saharan, and Khoisan families), Australia (in most subgroups of the Australian family), and New Guinea (in many branches of the Indo-Pacific family), as well as in many more language families from other continents, such as Afroasiatic, Turkic, Mongolic, Tungusic, Uralo-Yukaghiric, Japonic, Burushaski, Sino-Tibetan, Yeniseian, Dravidian, Eskimo, and Na-Dene, as well as probably, under a phonetically decayed form, Indo-European (Bancel & Matthey de l'Etang 2002; see Map 1).



Map 1. The global etymology *kaka* 'mother's brother, spouse's father, grandfather, elder brother' (sample data)

Languages are grouped in phyla, themselves arranged in columns according to their approximate respective location on the planisphere. Phylum names (e.g. DENE-CAUCASIAN) appear in capitals above or below each column, followed by the most likely original form and the kinship positions it most likely referred to; in each row, the language name (e.g. Zuni) is followed by the vernacular word in italics (e.g. *aga*), then by the abbreviated main meaning of the word (Fa 'father', Mo 'mother', Br 'brother', Zi 'sister', Sib 'sibling', So 'son', Sp 'spouse', Gd 'grand', Pt 'parent', Ch 'child', e 'elder', y 'younger', MoBr 'mother's brother', etc.); in Proto-Austronesian and Austronesian languages, Sib+ glosses words referring to an elder sibling of opposite sex to ego (elder brother of a female, elder sister of a male)

On the basis of the data from some 700 languages we had first investigated, we also determined that the focal etymological meaning of *kaka* was 'mother's brother' rather than 'uncle', followed by the less widespread meanings 'grandfather' and 'brother' (Matthey de l'Etang & Bancel 2002). We also suggested that kinship appellatives might indeed be much older than Proto-Sapiens, and that their simple phonetic form and specific use as calls by babies might have played a crucial role in the emergence of articulate language (Bancel & Matthey de l'Etang 2002).

Further work, relying on a growing database of kinship terminologies (now comprising over 2,200 languages), led us to develop our theories about both the Proto-Sapiens origin of kinship appellatives (Bancel et al. 2010; Matthey de l'Etang & Bancel 2005, 2008, in preparation; Matthey de l'Etang et al. 2010) and their role in the emergence of articulate language (Bancel & Matthey de l'Etang 2005, 2008, 2010).

1.2 Trask and the historical emergence of *papa/mama* words

Both Ruhlen's and our theses, however, were soon opposed by a comparative linguist, the late Larry Trask.¹ To defend Murdock's and Jakobson's theory of the multiple, spontaneously convergent origins of *papa/mama* words, Trask (2004) reviewed the history of these words in various languages, and concluded in favor of their "endless re-creation and recycling" (p. 15). It was the first time, in over a century and a half, that an attempt was made to substantiate the traditional theory from a historical viewpoint. Indeed, Trask's work was useful in forcing us to descend from a global, essentially statistical, viewpoint to the level of individual languages and families, in order to show that these words, contrary to Trask's claims, are not innovations in any particular language, but have been preserved throughout the histories of their respective families. Expanding on a previous answer (Matthey de l'Etang & Bancel 2008), our main task hereafter will be to show, with a wealth of comparative data, that Trask's study is flawed by fundamental fallacies, that none of his examples is an innovation, and that all of them are, instead, words that have been preserved over millennia with little or no change.

1.2.1 *Inherited papa/mama words in Indo-European languages*

By a radical misinterpretation, Trask confuses *papa/mama* with *father/mother* words. All his examples of lost or decayed *papa/mama* words are in fact *father/mother* words,

1. Trask did not quote our work or Ruhlen's, but there is little doubt that his study was intended as an answer to it, as it was published two years after our first papers had appeared (Bancel & Matthey de l'Etang 2002; Matthey de l'Etang & Bancel 2002) in the comparative linguistics journal *Mother Tongue*, of which Trask was an assiduous reader and contributor, always to defend the traditional view that no trace of common linguistic inheritance older than a few millennia should be taken seriously.

normal words of the standard adult lexicon, used to refer to any parent rather than to address one's own. Let us begin with the Indo-European family,² from which he draws numerous examples of “innovated” *papa/mama* words.

The Proto-Indo-European (PIE) words **patēr* ‘father’ and **matēr* ‘mother’ are, as Trask (2004, p. 12) himself says,

mama/papa words which have acquired a suffix *-ter* Already these words were being treated like other words in the language. Since PIE, the original words for ‘mother’ and ‘father’, where they have survived at all, have undergone the usual changes in pronunciation in the languages possessing them,

like Swedish *far* and *mor*, French *père* and *mère*, or Irish *athair* (phonetically [ahir]) and *mathair* [ma:hir]. So Trask concludes:

It is scarcely likely that anyone would recognize [ahir] as a *mama/papa* word, but in origin it definitely is. The *mama/papa* words are in no way resistant to the process of linguistic change, including regular changes in pronunciation. Nor are they resistant to loss. (Trask 2004, p. 12)

There is not the least doubt that PIE **patēr* and **matēr*, evidently derived from preexisting **pa(pa)* and **ma(ma)*, “in origin definitely” were *papa/mama* words.

But in origin only. Already in PIE, **patēr* and **matēr* were no longer *papa/mama* words – simple reduplicative words mimicking the babbling of babies and used to address one's own parents. Instead, they had become *father/mother* words – ordinary words of the PIE lexicon, used to refer to anyone's parents, as are all their derivatives in modern languages: English *father* and *mother*, German *Vater* and *Mutter*, Swedish *far* and *mor*, Icelandic *faðir* and *móðir*, French *père* and *mère*, Spanish and Italian *padre* and *madre*, Occitan *paire* and *maire*, Irish *athair* and *mathair*, Greek *patéras* and *mitéra*, Armenian *hayr* and *mayr*, Persian *padar* and *mādar*, Ossetic *fyd* and *mad*, and hundreds of others. Word replacement and phonetic change had to – and obviously did – apply normally to these normal words of the adult lexicon.

And **patēr* and **matēr* certainly were not the first words of PIE-speaking children some 7,000 years ago, any more than *father* and *mother* are the first words of English children today, or *père* and *mère* those of French children.

2. The Indo-European language family, whose discovery in the end of the 18th century and further exploration in the 19th gave birth to linguistic science, comprises most groups of languages spoken in Europe today (Celtic, Italic, Germanic, Baltic, Slavic, Albanian, Hellenic, Armenian), as well as the huge Indo-Iranian group, itself divided in three subgroups (Indic, Nuristani, and Iranian); it also includes two extinct groups, Anatolian and Tocharian. The reader unfamiliar with language classification will find members of each group listed in Appendices A to C and G, with examples of common words.

Moreover, Trask does not document a single *papa/mama* word known to be lacking in a given stage of a language's history, which appeared in a subsequent stage. He merely assumes that, in every language where *papa/mama* and *father/mother* words coexist, the former must be more recent than the latter. And in doing so, he often goes against their known etymology.

As we will see in detail below, all of his "new" *papa/mama* words have been inherited from the most ancient stages of their respective language families. When PIE **patēr* and **matēr* were derived as reference terms from Pre-PIE **pa(pa)* and **ma(ma)*, the more ancient forms did not disappear. **Pa(pa)* and **ma(ma)* must have been kept in parallel use as appellatives, just as in English *father* coexists with *dad*, and *mother* with *mom*. The reader is referred to Appendix A, which displays the etymological series supporting PIE **ma(ma)* 'mother, mom' in the Tower of Babel³ Indo-European database (Nikolayev 2007), completed by other data. From Prakrit *māmikā* 'mother', Classical Greek *mā* (*gā*) '(Earth) Mother' and Latin *mamma* 'mommy' to Punjabi *mā* ~ *māu* ~ *māi* ~ *māmmī* 'mother', Persian *mām* 'mom', Armenian *mam* 'grandmother', Modern Greek *mama* 'mom', Ukrainian *mama* 'mom', Latvian *māma* 'mom', Faeroese *mamma* 'mom', Sutsilvan Rumantsch *moma* 'mom', French *maman* 'mom', *mamie* ~ *mémé* 'granny', Breton *mam* 'mother', or Gheg Albanian *mame* 'mother', more than a hundred languages from the vast majority of IE subgroups unambiguously establish the PIE antiquity of this word.

Nikolayev (2007) does not posit a PIE root **pa* or **papa*. Its existence, however, cannot be doubted given the comparative data of Appendix B, which provides some 170 *papa* words from well over a hundred IE languages, from Palaic *pāpa* 'father', Prakrit *bappa* 'father', Khwarezmian *papa* 'father', Classical Greek *pappa* 'dad', *pappous* 'grandfather' or Latin *pappa* 'dad' and *pappus* 'grandfather', to Marathi *bāp* 'father', Kāmv'iri *vov* 'grandfather', Farsi *bābā* 'father, grandfather', Armenian *pap* 'granddad', Modern Pontic Greek *papa* 'dad', Latvian *paps* 'dad', Danish *papa* 'dad', or Occitan *papā* 'dad' and *papet* 'grandfather, granddad'.

The Proto-Indo-European descent of these words eliminates many of Trask's "innovations": Greek *mama*, Icelandic *mamma* and *pabbi*, Norwegian *mamma* and *pappa*, French *maman* and *papa*, Italian *mamma* and *babbo*, Polish *mama*, Bengali *ma* and *baba*, Hindi *baba* or *bap*, Persian *mām* and *baba*, Latvian *mama* and *paps* (Trask

3. The Tower of Babel Project (<http://starling.rinet.ru/>) brings together the Russian State University of the Humanities (Moscow, Russia), the Moscow Jewish University (Russia), the Russian Academy of Sciences, the Santa Fe Institute (New Mexico), the City University of Hong Kong (China), and the Leiden University (The Netherlands). It provides free access to etymological databases for numerous language families, compiled by some of the best specialists worldwide. In our etymological lists, unreferenced data not drawn from Nikolayev (2007) may be found in easily accessible standard dictionaries.

2004, pp. 13–14) all directly derive from PIE appellatives **mama* and **papa*, which themselves must be even older than PIE, since in PIE times their derivatives **patēr* ‘father’ and **matēr* ‘mother’ were already well established.

Two double examples, jointly presented by Trask to illustrate the converging process of innovation in kinship appellatives, are worth special consideration:

The ancestral PIE words [**matēr* and **patēr*] have been completely lost in a number of the daughter languages, lost and replaced by other words. Two of those languages are Romanian and Welsh [...]:

	‘mother’	‘father’
Romanian	<i>mama</i>	<i>tata</i>
Welsh	<i>mam</i>	<i>tad</i>

But look at the words which have replaced the lost older ones! The newer words which have replaced the older ones are themselves *mama/papa* words. According to the Proto-World account, ... [t]he *mama/papa* words are supposed to be no more than ancient survivals, and they can’t do anything except survive for a while longer or disappear. They absolutely can’t *reappear* in languages which have lost them. But they do. And they do it all the time. (Trask 2004, p. 12)

Reappear all the time? Romanian *mămă* certainly did not (re)appear out of the blue, nor did Welsh *mam*. The data in Appendix A establish that they were inherited from Latin *mamma* and Proto-Celtic **mama*, respectively, and that both ultimately derive from PIE **mama*. They are exactly the “ancient survivals” Trask does not want to see in them. And this has long been known to etymologists (Romanian: Meyer-Lübke 1911; Academia Română 1998; Welsh: Charles-Edwards 1993, p. 169).

But what about Romanian *tătă* ‘father, dad’, and Welsh *tad* ‘father’? Could they be “newer words which have replaced the older ones”? They could not. Romanian *tătă* has been known for a century to derive from Latin *tāta* ‘dad’ (Meyer-Lübke 1911; Ciorănescu 1958–66; Academia Română 1998). According to Charles-Edwards (1993, p. 169), Welsh *tad* goes “back at least to the Romano-British period” (43 CE to early 5th century), as it is found in all the ancient stages of the Brythonic group of Celtic (Old Cornish, Middle Welsh, and Middle Breton). And Old Irish (a language belonging to the Goidelic group) *data* ‘foster father’ shows that the word must be of Proto-Celtic origin.

But their antiquity in their respective language groups – Romance and Celtic – is not the end of their story. Both words belong to the PIE etymology **tata* ‘dad, father’ reported in Appendix C, again based on Nikolayev (2007) and completed with data from various sources. Once more, from Hieroglyphic Luwian *tati(a)-* ‘father’, Vedic Sanskrit *tatā* ‘father’, Old Avestan *tā* ‘father’, Classical Greek *tatā* ‘daddy’, or Latin *tāta* ‘dad’ to Kām̐v’iri *tot* ‘father’, Roshani *taat* ‘father’, Czech *táta* ‘father, dad’, Latvian *tēte* ‘dad’, Romanian *tătă* ‘father, dad’, Breton *tad* ‘father’, or Albanian *tatë* ‘father’, both

ancient and modern data from most subgroups abundantly testify to its inheritance from the earliest PIE stages.

To sum up, Trask's claim that Romanian *tătă* and Welsh *tad* are words that have recently (re)appeared is, again, contrary to obvious etymological and comparative facts.

1.2.2 *Inherited papa/mama words in Dravidian and Turkic languages*

Let us also consider two non-Indo-European examples cited by Trask, in Tamil and Turkish, respectively. In Trask's (2004, p. 14) view, the "informal" Tamil word *appaa* 'dad' is newer than the "formal" *takappaṇ* 'father'. But it simply cannot be. The honorific *takappaṇ* is a compound formed from *tak*, an adjective form of verb *taku* 'to be excellent', and *appaṇ* 'father' (Emeneau 1953, p. 342, 10). And Tamil *appaṇ* 'father' is itself a suffixed derivative of *appaa* 'dad', just as PIE **patēr* was a suffixed form of **pa-*, as revealed by the comparative data in Appendix D, drawn from the classical etymological dictionary of Dravidian.

The earliest trace of Tamil *appaa* is found in a 3rd-century CE inscription, used as a masculine honorific suffix (Mahadevan 2003, p. 609), as in Modern Kannada, Tulu, and Telugu. And *appaa* evidently derives from Proto-Dravidian.

With regard to Turkic, as Trask himself says, the inherited word for 'father' is *ata*, and this word is "still the everyday word in most Turkic languages." But, in Modern Turkish,

the word *ata* [has become] an elevated word meaning 'forefather, ancestor', [and] the everyday word for 'father' is now *baba*. This, of course, is another *mama/papa* word, and it used to be the Turkish word for 'daddy', but now it is the ordinary word for 'father', and 'daddy' must now be expressed by adding a diminutive suffix, producing *babacık*.
(Trask 2004, p. 13)

To Trask, this succession of a meaning shift, a replacement, and a suffixation illustrates the idea that nursery words change ceaselessly. Proto-Turkic **ata* 'father' is indeed reflected in many ancient and modern Turkic languages, from Old Uighur *ata* to Sary-Yughur *ata* through Tuvin *āda*, Azeri *ata*, and Khakassian *ada*, all meaning 'father' (Appendix E1). After 1,300 years, most of these terms remain identical to Proto-Turkic. In Turkic languages, preservation of **ata* has been the rule. Furthermore, its meaning shift to 'forefather, ancestor' in Modern Turkish is quite a common one. One's father is one's closest male ancestor, and in nearly all languages words meaning 'father' may also refer to other male ascendants, or even brothers and male descendants. This was the source of a vast majority of their semantic changes, which are merely expansions or retractions of their scope within the narrow field of kinship relationships, mostly within the same gender.

For its part, the diminutive *babacığım* 'daddy' (rather than *babacık*, which is not found in a single Turkish dictionary) does not replace *baba* (found in all Turkish

dictionaries with the meaning ‘dad’) in Turkish children’s first words nor in their parents’ baby talk, any more than English *daddy* replaces *dad*, Italian *babbino* replaces *babbo*, or French *papoune* or *papounet* replaces *papa*. They are affectionate diminutives, and may continue to coexist for centuries with their respective root words *baba*, *dad*, *babbo*, and *papa*, or perhaps enter the standard language with a new meaning. But *baba*, *dad*, *babbo*, and *papa* will remain, because babies need them to learn to speak, and parents to teach children, as will be explained in Section 3 below.

Finally, as for Turkish *baba* itself, far from being new, it was borrowed from Persian (Nişanyan 2001) after the Türks invaded the Persian Empire, a borrowing certainly facilitated by the existence in Turkish of another old Turkic word, *aba* ‘father, ancestor’, preserved in many Turkic languages (Appendix E2).

Borrowed? Yes, *papa/mama* words may be borrowed, and indeed they are probably more frequently borrowed than any other words in the basic lexicon. We have already met Greek *baba*, borrowed from Turkish – which had previously borrowed it from Persian. Albanian *baba* was also borrowed from Turkish during the Ottoman domination over the Balkans. English *dad*, an isolated form in the Germanic group of Indo-European, whose other members all have *papa* forms (Appendix B), was likely borrowed from Brythonic Celtic, where *tad* ~ *tat* forms are general (Appendix C), when the Anglo-Saxons invaded Great Britain. It is also likely that Romanian *tătă* ‘father, dad’, a descendant of Latin *tāta* ‘dad’, which even replaced in Romanian the outcome of Latin *pătēr* ‘father’, was helped to survive – and thrive – by the forms *tata* ‘father, dad’, which are general in the surrounding Slavic languages from which Romanian borrowed thousands of other words, while many other Romance languages lost Latin *tāta* and preserved *pappa* instead.

But borrowing is not an innovation, in the sense of a newly created word. A borrowed word has a history in the donor language, and the receiver language continues this history. In the case of Turkish *baba*, as we have seen, its Persian source derives from Proto-Indo-Iranian **baba*, itself from Pre-Proto-Indo-European **papa* (Appendix B). Brave new word.

1.2.3 *Inherited papa/mama words in Chinese languages*

The case of Chinese, not studied by Trask, also deserves consideration. In nearly all modern Chinese languages, from Mandarin to Cantonese, address terms used for one’s father and mother are *pa* ‘dad’ and *ma* ‘mom’, respectively (see Appendices F1 and F3). Only their tonal contours vary according to dialect. Both *pa* and *ma* have reduplicated variants, respectively *papa* and *mama*, felt to be more childish by speakers (Agnès Gaudu, personal communication).

In the Chinese Characters (Starostin 2006) and Modern Chinese Dialects (Wang 2004) Tower of Babel databases, modern *pa* ‘dad’ forms are assigned an etymology dating back to Preclassic Old Chinese *paʔ* ‘father’, implying very little variation over

some 3,500 years (Appendix F1). The ancient forms are apparently shared with the etymology of the referential word 'father' (Appendix F2), derived from Middle Chinese *pú*, a dialectal form attested since the Tang period (seventh to tenth centuries CE), evolved into Beijing, Jinan, or Xi'an *fu*³, Shuangfeng *ɣəu*³², Chaozhou *pe*²², Fuzhou *xu*³², Shanghai *vu*³², etc. (superscript numbers indicate tones).

In the phonetic form *paʔ* 'father' of Preclassic Chinese reconstructed by Starostin (2006), the final glottal stop -ʔ is essentially posited to explain the tonal evolutions in modern dialects, inspired from regular correspondences in other words. But in words like 'dad' and 'mom', the evolution of whose tonal pattern is highly likely to have been influenced by expressive intonational patterns, this particular final -ʔ does not need to ever have existed. And it surely did not, given that it is not present in even a single modern Chinese dialect.

Indeed, what happened in Chinese seems clear. From a Preclassic *pa*, a form *pwá* ~ *pwó* 'father, dad' appeared during the Han period and progressively specialized as a referential term, giving rise to Middle Chinese *pú* 'father', from which all the modern forms *fu* 'father' derive. Meanwhile, *pa* continued to be used as an address term in the spoken language and was transmitted without any change in all Chinese dialects. However, the pictogram that originally read *pa* received the phonetic reading of the reference term, showing that *pa* was originally used for both address and reference.

The identical (except for tones) *pa* forms of all modern dialects – known not only by ideograms, but by phonetic descriptions as well – prove that *pa* survived unchanged throughout the history of Chinese. And the Eastern Han and Postclassic *pwá* or *pwó* forms have been misattributed in Starostin's database – they may not be the phonetic ancestors of modern *pa* 'dad' forms but are forerunners in the evolution of modern *fu* 'father' forms.

A similar situation appears in the etymology of terms meaning 'mom' and 'mother', although the two terms may have begun to differentiate already in Preclassic Chinese (see Appendices F3 and F4). In ancient forms, again, the aspirated initial *m^h*- and the final glottal stop -ʔ are reconstructed on the basis of tonal developments in modern dialects. Thus, as in the case of *pa* 'dad', both are far from assured and indeed are superfluous, given their absence from *ma* words in all modern dialects and the expressive uses of appellatives.

Just as Proto-Indo-European speakers created **patēr* 'father' and **matēr* 'mother' from preexisting **pa(pa)* and **ma(ma)* words and continued to use them in parallel, Chinese speakers developed new reference (*father/mother*) words from the Old Chinese words *pa* and *ma* that were initially used for both reference and address. But speakers continued to use in parallel the original *pa/ma* forms as address terms. The new reference terms have strongly evolved in modern Chinese dialects, e.g. Wenzhou *voy*²² or Shuangfeng *ɣəu*³² 'father', or Chaozhou *bo*²¹ 'mother' – in which a non-nasal

consonant has even appeared – but in all dialects the address terms *pa* ‘dad’ and *ma* ‘mom’ have remained exactly the same as those used over 3,000 years ago.

1.3 Summary

- *Papa/mama* words are exempt from most phonetic evolutions, but may, on occasion, vary phonetically within the limits allowed by babbling as regards vowel quality and length as well as consonant gemination and the voiced/voiceless (or fortis/lenis) contrast, for example within the Germanic group, German *Papa*, Rhine Franconian *Päppe* ~ *Bäbbe*, Bavarian *Babba*, Faeroese *pápi*, Icelandic *pabbi* ‘dad’.
- Due to their once common use to address elders respectfully or youngsters affectionately, they may vary semantically, in general within the same gender, e.g. Sogdian *bābay* ‘father’, Yaghnobi (a modern descendant of Sogdian) *bobo* ‘grandfather’, but may also occasionally be recruited into morphological alternations, e.g. Bashkarik *mēm* ‘mother’s mother’, *mām* ‘mother’s father’, locally introducing etymological confusion.
- They may give rise to *father/mother* words and continue to coexist with them, for example, Pre-PIE **papa/*mama* having given rise to PIE **papa/*mama* and **patēr/*matēr*, or Old Chinese *pa/ma* having evolved into Mandarin *pa*³/*ma*¹¹ and *fu*³/*mu*².
- *Papa/mama* words may be borrowed, such as Modern Greek *baba* ‘dad’, borrowed from Turkish. Such borrowings are most of the time facilitated by similar preexisting words in the target language: Homeric *pappa* > Hellenistic *papa*, preserved in Modern Pontic, differed from Turkish *baba* only in consonant voicing. In turn, Old Turkish *aba* ‘father’ differed from Persian *baba*, borrowed into Turkish, only in partial versus full reduplication.
- Ancient languages possessed more *papa/mama* words than modern ones and used them extensively as terms of respect for elders. Certainly, PIE **papa* and **tata* were not exact synonyms; otherwise they would not have been preserved in so many of the descendant languages, always with different meanings in languages that preserve both. Their original semantic difference, which may have resided in their connotations rather than the persons they referred to, remains uncertain. Due to semantic overlap and the loss of importance of kinship relations – which used to be the very essence of the social organization in all hunter-gatherer societies, the only way of life of all human beings until some 10,000 years ago – some of these words have been lost in historical times, as was Latin *tāta* ‘dad’, lost in French, Occitan, Spanish, and Portuguese.
- A few such words, however, do randomly appear in the course of the history of individual languages, such as French *tata* ‘auntie’, a diminutive of *tante* ‘aunt’. Such

cases do not stem from babies' babbling but from adults' baby talk. Very importantly, they do not obey the distribution rule *Oral stops for males, nasal stops for females*. This rule was already observed in 75% of languages by Murdock (1959) in his survey of words meaning 'father' and 'mother' in 474 languages, and was confirmed by our own statistics bearing on 1,184 languages (Table 1; for a detailed analysis, see Bancel et al. 2010). And, given the massive preservation of original forms in most languages from all families, innovations may only represent a tiny minority of the countless *papa/mama* words worldwide.

Table 1. Most prominent meanings of *papa*, *kaka*, *nana*, and *mama*

Form Meaning	PAPA		KAKA		NANA		MAMA	
	Number of occ	Percent (tot. > 100)	Number of occ	Percent (tot. > 100)	Number of occ	Percent (tot. > 100)	Number of occ	Percent (tot. > 100)
F	288	59.2%	12	2.4%	38	6.2%	84	26.1%
F + FB	106		3		4		82	
FB	100	15.0%	59	9.2%	1	0.1%	7	1.1%
FZ	36	5.4%	10	1.6%	48	7.1%	31	4.9%
M	20	6.0%	14	3.0%	250	64.0%	232	43.1%
M + MZ	20		5		182		42	
MZ	8	1.2%	24	3.7%	48	7.1%	49	7.7%
MB	33	5.0%	221	34.5%	11	1.6%	105	16.5%
B+	100	15.0%	111	17.3%	28	4.1%	4	0.6%
Z+	27	4.1%	64	10.0%	59	8.7%	9	1.4%
Sib+	7	1.1%	32	5.0%	17	2.5%	9	1.4%
GdF	134	20.1%	86	13.4%	16	2.4%	15	2.4%
GdM	45	6.8%	61	9.5%	48	7.1%	52	8.2%
GdPt	15	2.3%	35	5.5%	4	0.6%	12	1.9%
GdPt + GdCh	42	6.3%	31	4.8%	4	0.6%	35	5.5%
GdCh	38	5.7%	28	4.4%	0	0.0%	6	0.9%
Ch	14	2.1%	3	0.5%	65	9.6%	35	5.5%
TOTAL out of 1,184 languages	1,033 cognates in 666 languages (56% of sample)		799 cognates in 641 languages (54% of sample)		823 cognates in 675 languages (57% of sample)		809 cognates in 635 languages (54% of sample)	

Figures calculated for 1,184 languages; percentages have been calculated with regard to the number of languages attesting one or more words in the series concerned. Not all kinship relations attested for each term are listed above: for each series, at least a dozen other relations are attested by a few items. As of August 2013, our database comprised more than 2,400 kinship terminologies, and percentages would not be very different. (Table from Bancel et al. 2010)

Both Murdock's semantic convergence rates and our own statistics have been calculated for *father/mother* words and *papa/mama* words taken together. The reason is, as we found in our own compilation of kinship terminologies, that while words meaning 'father' and 'mother' nearly always figure even in the shortest wordlists noted by field linguists or anthropologists, the corresponding appellatives are seldom noted – probably because of their perceived childish nature and their near identity in all languages, which result in a kind of disdain towards them. Since *father/mother* words are much less stable than *papa/mama* words, there is no doubt that the proportion of *papa/mama* appellative words complying with the oral/nasal distribution rule would be much higher than 75%, and probably above 90%.

However, before concluding that *papa/mama* words must share a common origin, we have to address another possibility.

2. Chance resemblances?

The main argument opposed to etymologies linking languages at a greater remove than Indo-European or other relatively recent ancestor languages is that the comparative series they rely on might have arisen by chance. To the best of our knowledge, this argument has never been leveled at *papa/mama* words, and we might consider it discarded in advance by the wide acceptance of Murdock's and Jakobson's theory of their spontaneous convergence under the influence of babies' babbling. If chance might have led to this convergence, putting forward or accepting any other explanation would have been absurd.

Indeed, the true absurdity would be to consider that the massive global convergence of *papa/mama* words could have arisen by chance. The overwhelming majority of these words are traceable to the very origin of their respective language family, in which they have survived for millennia – in the case of Indo-European languages, for 6,000 to 8,000 years, according to the most likely estimations. How could they have spontaneously emerged in different families all over the world with convergent meanings and phonetic forms in a distant past, while they have been among the most conservative words in the last several millennia? In ancient languages as well, they had to have been inherited, even if some may have been borrowed in a minority of cases.

The primordial role of kinship in the social organization of all peoples before the appearance of agriculture – and undoubtedly for eons, as clear precursors of kinship relationships are found in our closest ape cousins, chimpanzees and bonobos (de Waal 1982; Fouts 1997) – excludes the possibility that they could be recent inventions. Their global distribution definitely excludes generalized, intercontinental borrowings, so that the only remaining explanation is that they have been transmitted over dozens of millennia from a common ancestor language. They may only have stemmed from a common, Proto-Sapiens origin – an idea which makes sense with regard to

both archeological and genetic data about the expansion of *Homo sapiens* from their African homeland some 50,000 to 100,000 years ago.

We could thus dispense with a detailed refutation of the chance hypothesis. Nevertheless, we will address it here in some detail, as we are convinced that deep-time linguistic comparison has much more to tell us about the development of human language as far back as the beginning of our species' expansion, thus shedding light on a crucial period in *Homo sapiens*' history, with the dramatic acceleration of technical evolution and the appearance of food cooking and of personal and graphic ornaments. Many scholars think that these changes must be linked to an evolution of human language ability, with the most frequently mentioned candidate being the emergence of syntactic articulation. We have no doubt that the comparative-historical study of languages can help to understand this evolution, and we will illustrate this opinion at the end of this article (Section 5.1).

Through a detailed analysis of two tentative probabilistic refutations of deep-time etymologies, we will show that proving or disproving Proto-Sapiens etymological series by means of probabilities would demand calculations involving many parameters, some of which are not easily amenable, if at all, to numerical representation. It will also appear that the etymologies subjected to these treatments are beyond the point where a probabilistic assessment is necessary. Similarly, regular phonetic correspondences in low-level linguistic families are far beyond the level where chance might be involved and are with good reason regarded as indisputable proof of the common descent of the words they are found in, without having ever undergone any kind of mathematical assessment.

2.1 Inaccurate calculations

The probabilistic refutations of deep-time linguistic comparisons known to us fall into two categories. The first one is that of historical linguists unfamiliar with the basic principles of probabilities. For instance, the Indo-Europeanist Donald Ringe (2002), trying to show that Greenberg's (2000) Eurasiatic⁴ etymologies are due to chance resemblances, overlooks the fact that a probability is a ratio – that is, it describes the number of chances for a particular event to happen *out of* a total number of possible events, so that one has 1 chance *out of* 6 of getting an ace when throwing an ordinary die, but only 4 *out of* 52, or 1/13, when taking a card from a deck. This leads Ringe, in six dense pages, to multiply probabilities as he adds parameters that obviously shrink them – as if he had found that, when taking a card from each of four decks, there were

4. Eurasiatic is a macrofamily of languages discovered by Greenberg (2000–2001) encompassing the Indo-European, Uralo-Yukaghir, Altaic, Koreo-Nippo-Ainu, Gilyak, Chukchi-Kamchadal, and Eskimo-Aleut language families.

$4 \times 4 = 16$ “absolute” chances of getting 4 aces, instead of $(1/13)^4 = 1^4/13^4 = 1/28,651$, or 1 chance *out of* nearly 30,000. As a result, Ringe finds that Greenberg had “more than 35 quintillion” chances of discovering a first-person pronoun root **m-* common to 21 language groups from northern Eurasia.⁵ Out of how many possible outcomes, he does not mention, not realizing that 35 quintillion chances out of 3,500 quintillion would yield a tiny probability of 1%, or 0.01, while if there were 35 octillion possible outcomes, it would descend to a minuscule probability of 1 billionth, or $1/10^9$. No reliable conclusions can be drawn from such fanciful calculations.

2.2 Inaccurate comparative linguistics

The second category of erratic probabilities is due to scholars unfamiliar with comparative-historical linguistics performing apparently correct probabilistic calculations on irrelevant parameters. This is what the phonetician Louis-Jean Boë does with Bengtson and Ruhlen’s (1994) global – that is, Proto-Sapiens – etymologies, in a study whose successive versions (Boë et al. 2003; Boë 2004; Boë et al. 2006) do not show any real improvement in this regard.

2.2.1 Inaccuracy with regard to linguistic taxonomy

Knowing the proportion of languages that reflect an assumed original root seems important to ensure that the assumed cognate words are not random look-alikes: if you take a card from each of 52 decks, you may be nearly sure of getting at least one ace, and the greatest probability is that you will get four of them. How does this work with languages? Boë et al. (2003) count the total number of languages mentioned by Bengtson and Ruhlen (1994) in support of all their 27 Proto-Sapiens etymologies. They find 1,317 of them, and, assuming that this was the total number of languages investigated by Bengtson and Ruhlen, they relate to this total the average number of languages cited in support of each etymology. They find that each etymological series comprises an insufficient number of languages and families. But their count and its

5. This first-person root *m-* is represented in English by *me*, *my*, *mine*, and as a relic of the PIE conjugation system in *I am*. In an unpublished study, we have found that it survived as the first-person pronoun root in 99.6% of 494 Indo-European languages and dialects, from Icelandic *mi* to Assamese *mōk* through Portuguese *me*, Greek *mē*, Russian *menja* or Pashto *mā*, whose common descent from the PIE root **m-* is acknowledged by all Indo-Europeanists, including Ringe himself. Only two IE languages, Tocharian A and B, may have lost it. This stunning preservation, paralleled in most of the 20 other families alluded to by Ringe, from Turkic to Eskimo through Finno-Ugrian and Chukchi-Koryak, shows that chance has nothing to do with the presence of this pronominal root in 21 families, most of which also share a second-person root *t-* (English *thou*, *thee*, *thy*, *thine*) as well as some 70 other grammatical roots and hundreds of lexical roots (Greenberg 2000–2001).

alleged consequences are simply pointless. When introducing their Proto-Sapiens etymologies, Bengtson and Ruhlen warn that the potential descendant words they quote are only examples:

[S]ince the existence of these roots as characteristic features of the language families cited has already been established by other scholars, and is not for the most part in question, *we do not give the complete documentation for each family*, limiting ourselves in most instances to an indication of the range of semantic and phonological variation within the family. *The reader who wishes to see every relevant form for a given family should consult the sources cited.*

(Bengtson & Ruhlen 1994, p. 291; emphasis added)

Let us illustrate Boë et al.'s misinterpretation of Bengtson and Ruhlen's data – not for each of the 27 global etymologies, because that would take several books, nor even for a single one, but for a single family supporting a single etymology. In support of their Proto-Sapiens etymology *tik* 'finger, one', Bengtson and Ruhlen (1994, pp 322–323) give 184 reflexes from 165 languages (12.5% of the 1,317 languages they quote), including a mere 9 reflexes taken from only 6 Indo-European languages:

Indo-European: Proto-Indo-European **deik-* 'to show, to point', **dekm̥-* 'ten'; Italic: Latin *dig(-itus)* 'finger', *dic(-āre)* 'to say', **decem* 'ten'; Germanic: Proto-Germanic **taihwō* 'toe'; Old English *tahe* 'toe'; English *toe*; Old High German *zēha* 'toe, finger'.
(Bengtson & Ruhlen 1994, p. 322)

Does this sample exhaust what Bengtson and Ruhlen could have found in the Indo-European family? Well, not exactly. Appendix G1 displays the data mentioned by Nikolayev (2007) under the PIE etymology **deikē-* 'to show, to point', completed by Pokorný (1959), Lubotsky (no date), Turner (1962–1966) and standard dictionaries of various modern languages. While it is still far from exhaustive, it offers 170 derivatives of the Indo-European root **deikē-* 'to point, to show' in some 80 languages. As regards PIE **dekm̥-* 'ten', Appendix G2 lists 250 reflexes from 247 languages, drawn from the same sources plus the remarkable compilation of Rosenfelder (no date). The common descent of these words is assured by two centuries of Indo-Europeanist comparison and, as Bengtson and Ruhlen say, is "not for the most part in question."⁶

Thus, in the Indo-European family alone, over 400 possible reflexes of Proto-Sapiens *tik* add to the 9 examples given by Bengtson and Ruhlen. And Indo-European is but one of the 21 families displaying reflexes of Proto-Sapiens *tik* 'finger, one' in Bengtson and Ruhlen's series.

6. Only Classical Greek *dak-tulos* 'finger' and its direct Modern Greek descendant *ḍak-tilo* 'finger' are not recognized by Indo-Europeanists as related to the series (e.g. Chantraine 1968, pp. 249–250) because of their irregularity; we nevertheless think they do belong to it.

Boë et al.'s claim, based on language counts, that Bengtson and Ruhlen's etymologies are insufficiently supported, and thus likely to have resulted from chance resemblances, obviously falls far off the mark.

Now, should Bengtson and Ruhlen have published such huge lists for all families supporting each of their etymologies? From the viewpoint of reconstruction, no. The two PIE roots **deikē-* 'to point' and **dekṃ-* 'ten' rely on regular phonetic correspondences attested in innumerable other etymological series; hence their validity does not depend primarily on the number of reflexes but on the regularity in the detail of correspondences. No lists such as those in Appendix G have ever been published by any Indo-Europeanist, and this essentially underscores the vacuity of probabilistic calculations that do not take into account the fact that Proto-Indo-European is an ancestor language. With regard to the earlier history of a particular word, PIE represents *all* its descendant languages – those that preserved the word in question as well as those that lost it. If a word existed in PIE, the fact that it disappeared from 4, 40, or 400 descendant languages is irrelevant to the ancestry of this word *before* PIE, and Boë et al.'s method, beyond their misreading of Bengtson and Ruhlen's warning about the incompleteness of their examples, entirely misses this crucial point. Yet Bengtson and Ruhlen are quite explicit once again:

A common criticism is that, with around 5,000 languages to choose from, it cannot be too hard to find a word in some African language that is semantically and phonologically similar to, or even identical with, some word in an American Indian language. ... But this sort of mindless search is exactly the reverse of how the comparative method proceeds. The units we are comparing are *language families*, not individual languages. ... So instead of drawing our etymologies from thousands of languages, we are, rather, limited to [32] families, some of which have no more than a few hundred identifiable cognates. The pool of possibilities is thus greatly reduced, and accidental look-alikes will be few.

(Bengtson & Ruhlen 1994, pp. 279–281; emphasis in the original)

The inequality of languages and proto-languages with regard to their early history also affects contemporaneous languages: for instance, a reflex found in a language such as Basque or Burushaski, which by themselves constitute long-isolated language families, cannot be given the same etymological weight as a reflex found in one of the several hundred Romance or Germanic dialects. This evolutionary hierarchy is not easily reduced to figures – in particular with regard to disputed taxa, as is often the case of subgroupings within accepted families, and nearly always for remote macrofamilies and phyla: should Basque be given the weight of a completely isolated language, as if the Basques had independently discovered articulate language, or should it be considered a member of the Vasco-Caucasian macrofamily, or of Dené-Caucasian, a hotly disputed phylum whose huge range spans across northern Eurasia far into northwestern North America?

Still, from a probabilistic viewpoint, the number of languages in which a word from a proto-language did survive may not be entirely irrelevant to its earlier antiquity. The two lists in Appendix G tell us that the two PIE roots **deikē-* 'to point' and **dek̑-* 'ten' are among the words that have best resisted loss in the history of IE languages. In itself, this resistance shows that these words are able to survive over long periods of time, which is a strong a priori argument in favor of their ability to have survived over the times that preceded PIE as well. For this reason, Bengtson and Ruhlen might have published the detailed support of at least one of their etymologies.

But, whatever the amount of sources and data, we do not see how the taxonomic ranking of languages (i.e. the inequality between an ancestor language and its descendants, or between a long-isolated language and a dialect in a large family) could be taken into account in a statistical calculation. The recent achievements of cladistics, involving sophisticated probabilities, tend to show that it might perhaps be possible; but it would demand a serious collaboration between qualified statisticians and comparative linguists.

2.2.2 *Inaccuracy with regard to phonetic correspondences*

Boë et al.'s probabilistic assessment of the phonetic validity of Bengtson and Ruhlen's series is inaccurate as well. They total the different phonetic forms assumed by Bengtson and Ruhlen to descend from each root (Boë et al. 2003, p. 2707), and find it so large that, in their opinion, any correspondence would be allowed, and thus meaningless. The case of Proto-Sapiens *tik* 'finger', raised by Boë (2004) to illustrate Bengtson and Ruhlen's phonetic laxity, is again enlightening (Table 2).

According to Boë et al., the large number of different sounds reflecting each original sound (20 for *t*, 21 for *i* including diphthongs and loss, or even 26 if long vowels are counted separately, and 23 for *k* including loss) reveals Bengtson and Ruhlen's laxity in selecting their reflexes. And this laxity, of course, has severe probabilistic consequences.

But a glance at the *phonetic nature* of the sounds reflecting each sound in *t-i-k* shows that they form consistent sets, each defined by the region of the mouth where its member sounds are formed. Since a great majority of consonant evolutions preserve the original place of articulation, these sets thus encompass sounds most likely to evolve into one another.

All consonants reflecting the initial coronal consonant *t-* of *tik* are also coronals. Coronals constitute a class of sounds pronounced with the tip of the tongue raised close to or against the upper front teeth (interdentals, dentals) or just behind them (alveolars, post-alveolars). These consonants articulated in the same region of the mouth as *t* are known to derive from earlier *t*'s in numerous languages. Not a single labial such as *p*, *b*, *p'*, *β*, *f*, or *v*, nor a dorsal like *k*, *g*, *k'*, *γ*, *x*, or *χ*, which are extremely infrequent derivatives of a coronal consonant, appears in the series. Moreover, *t* itself occurs unchanged in 98 words out of 184, or 53.3%.

Table 2. Number of occurrences of each reflex sound in the 184 presumed cognates supporting Bengtson and Ruhlen's (1994) Proto-Sapiens series *tik* 'finger, one'

T	$\left[\begin{array}{cccccccccc} t & 98 & d & 23 & ts & 13 & s & 13 & \check{c} & 6 & d' & 4 & z & 4 & tl & 4 & th & 3 & ts' & 3 \\ t & 3 & t' & 2 & ts' & 1 & c & 1 & t & 1 & dl & 1 & s & 1 & \check{g} & 1 & tl' & 1 & r & 1 \end{array} \right]$															
I	$\left[\begin{array}{cccccccccc} i & 56 & (\bar{i} & 5) & e & 33 & (\bar{e} & 5) & a & 22 & (\bar{a} & 1) & o & 18 & (\bar{o} & 4) & u & 10 & (\bar{u} & 1) \\ \varepsilon & 2 & (\bar{\varepsilon} & 1) & i & 2 & \text{ɔ} & 2 & ə & 2 & \bar{o} & 1 & y & 1 & & & & & & & \\ ai & 3 & ia & 2 & ay & 2 & ei & 1 & yi & 1 & ai & 1 & ie & 1 & ea & 1 & oe & 1 & & & \\ \emptyset & (\text{zero}) & 5 & & & & & & & & & & & & & & & & & \end{array} \right]$										(monophthongs, with long vowels between parentheses)		(diphthongs)		(loss)	
K	$\left[\begin{array}{cccccccccc} k & 97 & g & 14 & \text{ɹ} & 10 & \eta & 7 & \check{c} & 6 & h & 5 & q & 4 & kk & 4 & k' & 4 & x & 3 & k^w & 2 \\ k & 2 & nk & 2 & c & 2 & y & 1 & kp & 1 & gb & 1 & kh & 1 & q^w & 1 & h'k & 1 & xk & 1 & jj & 1 \\ \emptyset & (\text{zero}) & 14 & & & & & & & & & & & & & & & & & & \end{array} \right]$										(loss)					

For each of the three sounds *t*, *i*, and *k*, the assumed reflex sounds have been counted. The relatively numerous sounds reflecting each of the original consonants *t* and *k* constitute mutually exclusive sets (with the sole exception of *č*, as it is a likely derivative of both *t* and *k*, particularly in the vicinity of an *i*). Vowels are much less stable in all languages, and the assumed reflexes of *i* cover the whole spectrum of vowel qualities; nevertheless, high front vowels close to *i* (*i*, *i*, *e*, *ε*) and diphthongs with an *i* or an *e* make up an overwhelming majority of the total (117 out of 184, or 63.6%).

If one then compares the set of sounds reflecting *t*- to that of sounds reflecting the final velar *-k*, one observes that they are mutually exclusive. Nearly all sounds reflecting *-k* are dorsal consonants like *k* itself. Dorsals constitute another broad class of sounds pronounced with the back of the tongue against or close to the hard or soft palates (palatals and velars, respectively) or the uvula (uvulars). All are known to reflect earlier *k*'s in numerous languages. The only exception is the postalveolar coronal *č* (with 6 occurrences reflecting *-k*, or 3.3%), which is a frequent outcome of a former *k* in the vicinity of an *i* or an *e* (e.g. Latin *civitatem* [kiwitate] 'city' > Italian *città* [čitta], or *centum* [kentu] 'hundred' > Italian *cento* [čento]). And *k* itself occurs unchanged in 97 words, or 52.7% of the total.

Obviously, the number of individual sounds reflecting each original consonant ought to be related to the number of phonemes that *do not* reflect this sound. And this relationship is easy to establish. No need to investigate the phonetic inventories of all the 1,317 languages counted by Boë et al. In their 27 etymologies, Bengtson and Ruhlen have used a clear principle: potential reflexes of a consonant essentially fall into six categories defined by their point of articulation: with the lips (labials), the tip of the tongue (coronals), and its back (dorsals). These articulatory features, which are among the most resistant in phonetic evolution, combine with the opposition oral-nasal, also very resistant to change. Thus, every consonant in a word has on average 1 chance out of 6 of falling into any of the six categories: oral labial, oral coronal, oral dorsal, nasal labial, nasal coronal, or nasal dorsal. For a two-consonant root like *tik*, there is $(1/6) \times (1/6) = 1/36$ chances that its two consonants will each fall into a particular category. And, in any given language, any two-consonant word root thus has 1 chance out

of 36 – or 0.028, a tiny probability indeed – that each of its consonants will fall within a particular category.

This parameter can be calculated correctly after all. And it shows that Bengtson and Ruhlen's alleged phonetic laxity is a strong constraint imposed on the discovery of potential reflexes.⁷

2.2.3 *Inaccuracy with regard to semantic correspondences*

Boë et al. finally find that Bengtson and Ruhlen are lax with regard to meanings as well. And this assessment appears to be just as accurate as that regarding sounds: the apparent variety is great, but the actual diversity is small. Let us again examine how the various meanings of the words reflecting Proto-Sapiens *tik* 'finger, one' quoted by Bengtson and Ruhlen are represented in their data (Table 3).

Table 3. Number of occurrences of each of the 30 different meanings in the presumed cognates supporting Bengtson and Ruhlen's (1994) Proto-Sapiens series *tik* 'finger, one'

'one'	67	'finger'	37	'hand'	23	arm	10	'ten'	9
'to show, point'	5	'toe'	5	'only'	5	'five'	4	'alone'	4
'index finger'	2	'middle finger'	2	'only one'	2	'fingernail'	2	'thing'	2
'first'	1	'to say'	1	'one by one'	1	'thumb'	1	'once'	1
'foot'	1	'with the fingers'	1	'in hand'	1	'to carry in hand'	1	'by ones'	1
'paw'	1	'single'	1	'forefinger'	1	'palm of hand'	1	'guy'	1

Total number of occurrences = 194 (> 184 because of a dozen words with two meanings).

Here again, 30 different meanings are represented in Bengtson and Ruhlen's series. But a glance at the number of occurrences of each meaning in their sample immediately shows that the two main meanings, namely 'finger' and 'one', which are closely linked together by the universal habit of counting on one's fingers, account for 104 of the 194 total meanings, or 53.6%.

The other, less-represented meanings should not be counted as weakening the numerous convergent words meaning 'finger' or 'one' – or Bengtson and Ruhlen could simply have not included them in their series in the first place, just as they did not include words meaning 'elephant' or 'carmagnole', even if they might have fit phonetically. Though coherent with the two basic meanings from a historical viewpoint,

7. We did not take into account the fact that consonant devoicing is respected in 79.9% of sounds reflecting *t*- and in 73.9% of those reflecting *-k*, nor of the fact that 63.6% of vowels are close phonetic images of *-i*-; these non-exclusive features are more difficult to integrate, but may only have a further strong restrictive effect on the probability that the series might have emerged randomly.

words meaning ‘hand’, ‘five’, ‘once’, etc. represent a bonus, often powerful when they are known to descend from an original word with one of the two critical meanings in their low-level family.

Moreover, the validity of an etymological meaning does not depend only or even primarily on the number of attestations of each modern meaning reflecting it, but much more on the reconstruction of a semantic evolutionary process. The original meaning of a word may have survived in few or even none of its descendants, while derived meanings may have proliferated. Obviously, Bengtson and Ruhlen’s *tik* series would have been much weaker if each of the 30 different meanings in their 184-word sample had been represented by 6 or 7 words, distributed without any evolutionary logic over the 21 families where reflexes of *tik* are found, contrary to what may readily be observed in the sample of Indo-European reflexes of *tik* in Appendix G.

The probability of finding a root with an initial *t*- (or any other oral coronal) followed by a *-k* (or any other oral dorsal) with either of the two meanings ‘finger’ or ‘one’ is double of that of finding a phonetically fitting word with only one particular meaning. As a result, 1 language out of 18 (instead of 36) should display consonants from two particular sets in a word with one of the two meanings ‘finger’ or ‘one’ by the effect of chance. This probability of 1/18, or 0.056, is still low, and it should apply, following Boë’s method, to all 104 languages where words with one of these two meanings have been found. (But we have seen in Section 2.2.1 above that these 104 languages are far from being the only ones to take into account, and, moreover, that their number is not really relevant.)

The 90 words with other meanings should be given a higher probability, though certainly not of 1, depending on the number of evolutionary steps separating them from the original meaning and on the number of words likely to be reached at each step. But calculating their respective probabilities, for each word in each language, would require very long investigations, which are not necessary with Bengtson and Ruhlen’s etymologies. In the 21 families where they found it – out of their 32 low- or medium-level language families covering all existing languages – *tik* must have had ‘finger’ or ‘one’ as its etymological meaning in at least Niger-Congo, Nilo-Saharan, Afroasiatic, Uralic, Korean, Eskimo-Aleut, Yeniseian, Sino-Tibetan, Na-Dené, Miao-Yao, Daic, and Amerind, to which one can likely add Indo-European and Turkic. To retain only the most secure ones, there are 12 ancestral languages displaying a root meaning ‘finger’ or ‘one’ with an initial coronal and a final dorsal consonant, a phonetic configuration which should occur by chance in 1 language (or ancestor language) out of 18 – not in 12 out of 32. The actual presence of *tik*-type roots with secure meanings ‘finger’ or ‘one’ in 37.5% of the world’s language families is thus at least 6.8 times above the 5.6% chance level. And this gap between chance and facts could only be enhanced, though more modestly, by the 9 other families with less strong semantic correspondences.

In short, counting the number of different meanings reflecting an original meaning in order to assess the plausibility of an etymological series is, strictly speaking, meaningless. For each word reflecting the proposed root in a given language, the plausibility of its semantic derivation (if any) must be assessed in the light of related words in its family as well as in closely related families. In Appendix G1, we can see that the PIE root **deikē-* 'to point' has descendants endowed with verb meanings as different as 'to point out, to show, to exhibit, to confess, to say, to teach, to accuse, to manifest, to give a sign' and others, plus nouns as disparate as 'direction, region, part, earth, world, camping ground, country, village, cultivated field, side, span, hand span, amazement, finger, toe, accusation, sign, example, token, dedicace, discourse', and 'judge', totaling 31 different meanings (and more could be added). Is PIE **deikē-* disqualified by this variety? Certainly not, because the variety is only superficial, and in each Indo-European subgroup meanings are organized into apparent logical evolutionary chains. This evolutionary logic cannot be adequately accounted for by a statistical model.

2.2.4 Summary

- The negative conclusions of the probabilistic calculations we have examined (Boë 2004; Boë et al. 2003, 2006; Ringe 2002) cannot be regarded as valid.
- Although it seems relatively easy to take into account the degree of phonetic validity of assumed reflex words, it is very difficult to reduce to figures the differences in taxonomic level between languages (the greater etymological weight of, e.g., Proto-Indo-European against any of its descendants, or of Basque against Gascon), or in logically derived meanings in a linguistic lineage versus meanings picked up at random without regard to semantic evolutionary logic (e.g. the logical validity of deriving 'toe' from 'finger', against the invalidity of directly deriving 'toe' from 'to point'). More work will be necessary to perhaps achieve a satisfactory assessment of etymological series by mathematical means.
- A point that is relatively difficult to conceive and understand is how multilateral etymological series differ from phonetically regular etymological series in lower-level language families. The latter (like those shown in Appendices G1 and G2, PIE **deikē-* 'to point, to show' and **dekm-* 'ten', respectively) aim to trace with certainty the descent of a root in all the descendant languages. The most powerful tool to ascertain that words from different languages belong to such etymologies is regular phonetic correspondences, which may practically eliminate any doubt that a particular word displaying them belongs to a given series, without any probabilistic assessment being needed – not because there is any magic in regular correspondences, but because they link together dozens of word series by their constituent sounds in metaserries whose appearance by chance would obviously have been highly improbable, just as no calculation is needed to realize that, say, getting 200 aces of hearts when taking a card at random from each of 200 decks is a near impossibility.

Multilateral series, in turn, rely on phonetic correspondences that are often not demonstrably regular in the state of our knowledge; in other words, they are not found again and again across different word series. But the phonetic nature of these correspondences otherwise complies, within each series, with evolutionary rules that have been discovered, over the last two centuries, in low-level families thanks to regular correspondences. As we have seen in Section 2.2.2, these rules impose a strong constraint on the discovery of potential cognates. This constraint is, however, weaker than that posed by regular correspondences themselves and does not warrant that each particular word included in a series really belongs to it; nevertheless, if many words in a series repeatedly satisfy this constraint, the likelihood that the entire series has appeared at random quickly drops. Consequently, a multilateral series warrants the authenticity of a root in a proto-language, while none of its assumed descendants may be considered to descend from it with perfect certainty – even if, taken collectively, most of them must descend from it.

This apparent paradox was expressed by Bengtson and Ruhlen:

We do not harbor illusions ... that every etymological connection we propose will be found, ultimately, to be correct, but we do believe that the removal of such errors as may exist in these global etymologies will not seriously affect the basic hypothesis, which does not depend on any specific link for its validity.

(Bengtson & Ruhlen 1994, p. 292)

What replaces regular phonetic correspondences in multilateral series is the number of families involved in them, and the recurrence of series within a particular group of families, such as that of **m-* ‘first person’ and **t-* ‘second person’, which are (together with many others) particular to the group of families Greenberg (2000–2001) calls Eurasiatic.

- Many etymologies presented by Greenberg and Bengtson and Ruhlen, including the ones discussed above, are so massively supported that no probabilistic calculation is needed. Just like *papa/mama* words or, for that matter, reconstructions supported by regular sound correspondences, they are far beyond the point where sophisticated tools might be necessary.
- However, accurate probabilities might be useful to uncover other, less well-preserved roots, to assess disputed taxa, and more generally to enlarge our comparative knowledge of remote language families. One can only encourage both statisticians and comparatists to continue to address this difficult problem in a constructive spirit.
- If *papa/mama* words have managed to last for several dozen millennia, why could not some other words have resisted as well? And perhaps not so few of them – after all, Bengtson and Ruhlen’s 27 Proto-Sapiens etymologies result from the efforts of two scholars, while hundreds of Indo-Europeanists have worked over the two last centuries on a few dozen closely related languages to unearth some 2,500 PIE roots.

3. Why kinship appellatives do not change: Children babbling, parents choosing

Let us now examine two lines of evidence from different fields of the study of language, which converge with our own to support the hypothesis that *papa/mama* words must have played a crucial role in the early appearance of articulate speech.

Papa/mama words have been preserved over the whole history of language families with a written tradition, as documented in Section 1 for a number of such families, and comparison within language families with no written record shows that such is the case for them as well (Matthey de l'Etang & Bancel 2012). Why is it, then, that they are not – or, at least, very infrequently – subject to phonetic change and word replacement, as all other words are? The reason is simple and compelling, and every parent who has raised a child who developed normal speech knows it, but this common experience has percolated into the domain of scientific knowledge only recently and without attracting much attention. *Papa/mama* words are crucial for babies to learn and for parents to teach babies to speak. The actual mode of their transmission has been explained by the language acquisition specialist John Locke (1990), and it is a nice piece of collaboration between parents and children.

Around the age of 6 to 9 months, on average, all babies enter the babbling stage of language acquisition. Canonical babbling consists of repetitive *bababa*, *papapa*, *mamama*, *dadada*, *tatata*, and *nanana* syllables, made up of plain labial or coronal nasal or oral stops, plus an open vowel (Oller 1980). It has long been recognized that these syllables are the first to be mastered by children because they are the easiest, due to a range of constraints (Westermarck 1891; Jakobson 1960; MacNeilage & Davis 1990; MacNeilage 2008).

Among these sequences, parents “recognize” those corresponding to a word in their language and reinforce them – notably by repeating them in their standard form while pointing a finger at the parent concerned – while they leave unreinforced other sequences that do not match with a word in their language, and which the child will thus progressively abandon.

This was Locke's (1990) great discovery, which definitively falsifies the theory of the spontaneous emergence of these words. Or, rather, it falsifies the theory that babies invent them alone. Children spontaneously provide a range of syllabic frameworks, and parents rectify some of them into the canonical forms of the corresponding words in their language: English parents reinforce *dadada* and *mamama* into *dad* and *mom* (or *mum*), respectively; French parents reinforce *papapa* and *mamama* into *papa* and *maman*, respectively; Turkish parents reinforce *bababa* and *nanana* into *baba* and *anne* (a word related to Proto-Turkic *ana* ‘mother’, also inherited in Turkish; see Appendix E3); and so on. It would never occur to a monolingual English mother to induce her daughter to call her *anne* (even if her own given name is Ann or Annie),

nor to monolingual French parents to recognize in their baby's babbling of *dadada* a word meaning 'dad' and to reinforce it.

This crucial way of transmitting *papa/mama* words explains why English children consistently learn *dad* and *mom*, French children *papa* and *maman*, Turkish children *baba* and *anne*, Hindi-speaking children *bap* and *mā*, and so on. Each of these words belongs to the lexicon of a particular language. Children provide the initial spontaneous syllabic framework; the exact phonetic form and meaning of each word are taught by parents. This fact, which clearly implies lexical inheritance rather than innovation, was elusively recognized by Jakobson (1960) in the paper in which he paradoxically argued for spontaneous innovations instead of common descent:

[C]hildren, being prompted and instigated by the extant nursery words, gradually turn the nasal interjection into a parental term and adapt its expressive make-up to their regular phonemic pattern.

This "prompting and instigation by extant nursery words" discreetly acknowledges the fact that parents reinforce their child's babbling and shape it into already existing words. And behind this teaching stands an uninterrupted transmission from generation to generation.

This specific mode of transmission also explains why these words change so rarely. When a language is in the process of undergoing a phonetic change that should change their form – for instance when stops between vowels change to fricatives (a very common type of change), so that *baba*, *papa*, *dada*, and *tata* should become *bava*, *paŋa*, *daʒa*, and *taʒa*, respectively –, the bio- and neuromechanical constraints bearing on babies who are learning to speak at that particular time are most of the time stronger. Babies do not master fricatives and continue to say *baba*, *papa*, etc. preventing the change from applying to the word in question; parents recognize the form *baba* or *papa* they have heard since their own childhood and reinforce it rather than the modified form, which in any case exceeds the baby's articulatory capacities.

As a result, these words are transmitted from one generation to another without change, and are unlikely to be lost, since the same spontaneous syllabic frameworks reappear every time another child reaches the age of 6 to 9 months and begins babbling – a phenomenon which must have occurred regularly in all human groups that have survived long enough for us to know something of their language, and thus have covered nearly all periods of phonetic change in all languages.

Finally, *papa/mama* words are crucial in another aspect of language transmission. In children's first utterances, they function no differently from animal communication. They have been dubbed *holophrastic words* ("a whole phrase in one word"; see De Laguna 1927), because they seem to convey information that should be rendered in adult language by a complex sentence. Brigaudiot and Danon-Boileau (2002), in a

section entitled “Les premiers *maman*, holophrases ou énoncés à un terme” [The first *maman*, holophrases or single-term utterances], quote a century-old analysis:

Childish *mama*, translated into advanced speech, does not mean ‘mother’ but rather a sentence such as ‘Mama, come here’, ‘Mama, give me...’, or ‘Mama, put me in the chair’, or ‘Mama, help me’. (Stern & Stern 1907)

These holophrases are similar to the calls of young animals “holophrastically” calling their mothers, except that the human baby’s call, contrary to those of all other animals, is phonetically articulate: it consists of vowels and consonants arranged into syllables. But *papa/mama* words do not remain mere calls for long. Within a few weeks or months, reinforcement by elders, together with the recurrence of the association, in the parents’ speech, of one particular reinforced sound sequence with the presence of the mother, and of another one with the father, induces the child to establish a link between each of these sequences and a particular being in the outside world. And this association is crucial, since it opens the door of symbolic meaning for the child.

In this way, too, parental appellatives play a unique role in the transmission of language. And it must have been so for untold ages.

4. Back to Proto-Human: The Frame, then Content hypothesis

Papa/mama words have survived – or, rather, their continuous transmission and preservation was necessary to our ancestors – over the last 2,000 to 10,000 generations. During this period, they have been crucial for babies learning to speak – and their parents teaching them – in the nice collaborative effort described by Locke. Why should they not have been preserved over the 20,000 to 100,000 generations before that? We suggested long ago that kinship appellatives might have been among the very first phonetically articulate sounds (Bancel & Matthey de l'Etang 2002), no doubt a long time before Proto-Sapiens was spoken.

At that time, we were not aware of Peter MacNeilage’s (2008; see also MacNeilage, this volume) “Frame, then Content” phonetic theory of the origin of speech, first presented with respect to modern babies by MacNeilage and Davis (1990). This theory holds that *papa/mama* words must be the first sound sequences mastered by a human mouth, for compelling phonetic reasons discovered through the observation of language acquisition. To understand these reasons, one has to recall that all humans speaking a language, whatever their individual differences, are true virtuosos – just like all falcons are nonesuch sky-divers, or all whales are outstanding apnea sea-divers, as a result of major selective pressures.

As explained by Lieberman (1992) – whose pioneering work (e.g. Lieberman 1975, 1985, 2000) opened the door to the study of language evolution, which had

remained barred for a century – speaking is the most difficult motor activity, because of the extreme speed and precision of the successive motions involved in the articulation of a speech sound string. According to MacNeilage (2008), about 40 different muscles are involved in the production of the various speech sounds, each performing a very different function such as controlling the pressure of the airflow breathed out of the lungs, or the tension of vocal cords, opening and closing the nasal airway, or giving the vocal tract a particular shape. Based on an average 15 of these muscles being involved in each particular sound, and a speech rate of 15 sounds per second, MacNeilage arrives at the fantastic number of 225 muscular actions per second in speech, or one every 5 milliseconds. Most of them must be effected with millimetric precision, and all must be tightly coordinated; otherwise the sounds produced are not those intended. Such defects in coordination do indeed happen and are a major source of phonetic evolution, showing that when speaking we are always at the extreme limit of our capacities, without even being aware of it.

On the auditory side, the high speed of some 15 to 25 units per second at which speech sounds are normally delivered is equally amazing. Hearers decode them easily, although it is often beyond the speed limit of 15 units per second beyond which other sounds merge into an undifferentiated buzz in the hearer's perception. And the brain areas and connections able to process this high-speed auditory flow can do so efficiently only after appropriate training – that is, learning the language. Just think how difficult it is, when you start learning a new language, merely to perceive the sounds you are not used to.

The extreme difficulty of both speaking and hearing an articulate language may be the reason why babies spontaneously start babbling in the second half of their first year. This universal behavior must rely at least partly on an innate trend, resulting from a heavy selective pressure exerted on humans to begin speaking at an early age, so they can gain the required fluency, again confirming that articulate speech has long been a major feature of the human ecological niche.

It also explains why babbling consists of plain stops and vowels in the simplest syllable sequences. Babbling typically reduplicates the most basic articulate syllables, namely consonant-vowel (CV), in CVCVCV... sequences. These syllable sequences using only two sounds are the easiest way to produce an articulate speech flow, as they require the synchronization of very few muscles into a repetitive, dual motor scheme, however long the syllable sequence may be. Moreover, MacNeilage and Davis (1990) found in early babbling an inertial pattern whereby the tongue stays in the same position for the vowel as it was in for the previous consonant, or the surrounding consonants in reduplicative babbling. Consequently, to produce *bababa*, an infant initially needs only a couple of mandible elevation muscles and a couple of mandible depression muscles. For *dædædæ*, she only needs to add the inferior genioglossus to move the tongue forward and up, and for *gogogo* only two or three muscles are added to the ones

used for mandibular oscillation (Peter MacNeilage, personal communication).⁸ From the neuromotor viewpoint, this is a huge simplification with regard to the requirements of adult speech.

The complete closure of stops also allows much more variation in the articulatory motions than for any other speech sound. No matter what the speed, strength, and precision of the closing motion to produce a stop may be – whichever way the airflow is closed and reopened, it will produce an acceptable approximation of the intended sound. In contrast, other consonants such as fricatives or glides demand millimetric precision in their execution, and any deviation from the intended target is likely to drastically modify the acoustic output.

Furthermore, as already noted by MacNeilage and Davis (1990), a babbling sequence essentially relies on motions that lower and raise the jaw – a motion over which voluntary control has been selected in the human lineage since our distant Gnathostomata ancestors, which appeared some 450 million years ago, acquired a mouth with a jaw.

The articulatory, motor, and syllabic robustness of consonants *p*, *b*, *m*, *t*, *d*, *n* is the reason why these speech sounds are the first ones children regularly master in the articulated syllable sequences *papapa*, *bababa*, *mamama*, etc. Of course, if one randomly “tries” one’s articulatory organs in order to make a sound, any human phoneme (and many other sounds) may result. However, when it comes to reproducing a sound, and – which is still more difficult – a sequence of two sounds at will, of course the easiest sounds and sequences must be the first to be mastered.

This is exactly what children do, and what humans learning to speak – either early in life or at an older age – must always have done. Both MacNeilage (2008) and we (Bancel & Matthey de l'Etang 2002, 2005) independently arrived at the conclusion that this rule must have been in force since the very beginning of articulate speech – not only in *Homo sapiens*, but in our more ancient human ancestors as well.

8. Our comparative data converge with MacNeilage and Davis’ (1990) finding concerning the detail of vowels in early babbling. According to them, children’s first velar consonants occur with a velarized vowel in sequences like *gogo* or *koko*. While compiling our *kaka* etymological series, we were soon struck by finding a high number of *koko* ~ *kuku* (or *gogo* ~ *gugu*) forms, sometimes even predominant over *kaka* forms, as in Nilotic or Southern Amerind, but also occurring sporadically in many other language groups. In contrast, *popo* ~ *bobo* and *toto* ~ *dodo* variants of *papa* ~ *baba* and *tata* ~ *dada* are extremely rare. We could not find any consistent semantic correlate of this variation in the vowel. MacNeilage and Davis’ finding regarding modern children may be regarded as confirmed by this globally frequent variant. Reciprocally, while it does not help to resolve the question of the antiquity of these *koko* ~ *kuku* variants, it provides an explanation for their existence.

5. By way of conclusion: The early steps towards articulate language

As we have seen, there are three lines of independent findings. The first finding is that of MacNeilage and ourselves, based on the phonetics of language acquisition, that *papa/mama* sound sequences are the obligatory first steps towards mastering articulate speech, and must have been so throughout human history. The second one is that of Locke, showing how children and parents cooperate in the transmission of *papa/mama* words; even if Locke himself does not consider the issue from an evolutionary perspective, there is no doubt that this mode of transmission is not recent in humans. And the third is Ruhlen's and our own finding, supported by data from thousands of languages worldwide, that most *papa/mama* words can only have been inherited from a common Proto-Sapiens language. All three lines of evidence converge on a scenario in which kinship appellatives must have early played a prominent role in the evolution of speech in humans and might even have been at its very origin.

Beyond this striking convergence, this scenario has other aspects adding to its evolutionary value. In particular, the initial acquisition by babies of phonetic articulation in their babbling stage through meaningless syllable sequences, some of which are then given a meaning by parents, seems to be a step towards the solution of a mystery that has barely been noted, much less explained, since research about language origins has burgeoned.

Words have to have been invented, however long this invention may have taken.⁹ But how? Both phonetic articulation and referential meaning are unprecedented in animal history, and both are too complicated to have been developed simultaneously. The first step towards the elucidation of their origin must therefore be to discover which appeared first. Babbling babies show us that phonetic articulation appears first in all contemporary individuals. And it must have been so originally as well, since speech is such a difficult activity that, if humans had found another way to convey referential meanings in the beginning, they certainly would not have gone to the trouble of trying to move their tongues and lips at an incredible speed from one incredibly weird position to another but would have stuck to the previously used means of expression and developed it further. Articulate speech must have been discovered by chance, as was the case with all biological evolutions before and after it, and in its simplest form – that of *mamama*, *papapa*, *bababa* sound sequences. It must also have been initially used to fulfill previously existing communicative, non-referential functions. Only later, probably much later, did its wonderful but highly demanding properties

9. The reluctance to deal with the emergence of words is most conspicuous in Kenneally's (2007) book *The First Word*. In spite of its title, this summary of the current state of research about language evolution does not even allude to the question.

allow for a very slow differentiation of sequences based on very few consonants. It opened the door to a functional differentiation, which, ultimately, led to the emergence of semantic reference.

The consonants in kinship appellatives already delineate a simple phonetic feature system, based on articulatory motions and the corresponding bundles of neuromotor commands, each of which must be called into play with different command bundles to produce different consonant sounds. Appellatives also constitute a simple semantic system, based on a few obvious semantic features, the first being the opposition between males and females. They thus offer a plausible path to the development of structured phonetic and semantic systems, whose interrelated features have made us the “symbolic species” (Deacon 1997, p. 87, Figure 3.3).

Finally, let us allude to the fact that kinship is another uniquely human trait, whose insertion in the humanization evolutionary process has hardly been discussed before, in spite of the many promising avenues it offers. Articulate language, this essentially social human ability, might not have developed without a reinforcement of social bonds, and kinship has long been the primary mode of human social organization. The antiquity of kinship is warranted by both the universality of kinship systems in all known human groups and the existence of precursors of kinship relationships in apes. Given the complexity of both language and kinship, it is only natural that they have coevolved, further enhancing the plausibility that the first symbolic meanings ever acquired by humans concerned kinship relations.

5.1 How else may Proto-Sapiens aid the study of language origins?

Finally, let us illustrate briefly how remote etymologies could shed light on other aspects of the evolution of language ability. Apart from *papa/mama* words, the most resistant words worldwide are first- and second-person pronouns (Dolgopolsky 1964; Pagel 2000). In all families, they display an incredible resistance, as compared to the average replacement rates of 13% to 18% per millennium that have been calculated for the 100 or 200 most basic words (body parts, natural elements, kinship relations, pronouns, basic verbs, etc.).

In an unpublished study bearing on 494 Indo-European languages and dialects, we have found that the PIE first-person **m-* and second-person **t-* have been lost, after 6,000 to 8,000 years, in an amazingly small number of descendant languages. First-person **m-* was lost by only two languages (0.4%), which amounts to a loss rate of 0.05% per millennium, granting **m-* a half-life of 1.38 million years.¹⁰ In turn, **t-* has

10. The calculation of the half-life of words was devised by Pagel (2000). It is not as reliable as its prototype in physics, where one observes the decay of a given quantity of an element

disappeared from seven languages (1.4%), which endows it with a loss rate per millennium of 0.18%, and a half-life of 385,000 years.

Personal pronouns from most language families display similarly minuscule loss rates. However, unlike *papa/mama* words, and contrary to what might be expected given this extraordinary longevity, there is *no global convergence* of phonetic forms and meanings in first- and second-person pronouns. We have studied (Bancel & Matthey de l'Etang 2008, 2010) the phonetic distribution of pronoun roots in shallow-time ancestral pronominal forms worldwide compiled by Ruhlen (1994b, pp. 252–260) – who, interestingly, did *not* discover any Proto-Sapiens first- or second-person pronoun root in spite of the pleasure he no doubt would have had in finding one. We have found that a majority of these pronoun roots are based on a handful of consonants, which, however, are distributed among the first and second persons in apparent disorder at the global level. A root *m-* may represent the first person singular in some phyla (like Eurasiatic or Niger-Congo), or the second person singular in others (like Amerind), and the same holds true of the other globally widespread pronominal root consonants *t-*, *n-*, *k-*, and *s-*, in spite of their monolithic coherence at the family-internal level.

Another salient aspect of the phonetic distribution of pronominal root consonants is the near absence of plain oral labial stops (*p-*, *b-*), with very few exceptions, and those few are often demonstrably secondary, such as *bi-* ~ *be-* ‘I (nominative)’ in Altaic languages.¹¹ While this global absence remains unexplained, its very existence must be considered as indicative of a relationship between all pronominal forms. Given that plain oral labial stops are among the most frequent consonants in the world's languages (Maddieson 1984, 1997) and are rather resistant to phonetic change, if first- and second-person pronouns had independent origins in many language families, a good number of them ought to be based on a root *p-* or *b-*.

We then thought that first- and second-person pronouns (and first- and second-person markers more generally) may have emerged only with the fluent use of syntactic articulation, and the necessity to quickly differentiate the speaker and the hearer in a complex sentence. In the stages that preceded the evolution of syntactic articulation in a broad sense – stringing words together – words were mostly used in isolation, and a great proportion of the speaker's intentions had to be inferred from the context. Words were, however, highly useful, as compared with no words at all, thanks to their

over time. With words, one may only observe their loss as the ancestral language splits into multiple descendants. It does, however, give a good indication of their relative stability.

11. The Altaic language family consists of Turkic (Turkish, Uzbek, Kazakh, etc.), Mongolic (Classical Mongolian, Khalkha, Buriat, etc.), and Tungusic (Manchu, Evenk, Nanai, etc.); Korean and Japonic, thought by Greenberg to be related to the former groups at a greater remove within Eurasiatic, are often included within Altaic by Nostraticists.

property of referring to objects or actions known to the speaker and the hearer. They gave the hearer an anchor to infer the rest, in a world where human activities and interests were much more restricted and predictable than in any society known to us today. But first- and second-person pronouns have the strange and unique property of shifting reference with the speech turn. One does not see how such words, deprived of the essential quality of words at that time of referring to a stable object, could ever have appeared, nor what use they might have had – a single-word sentence “me” or “thou” would have given little information to the hearer. When syntactic articulation first began, verbs may only have been “action or state words,” with no mark of tense, voice, person, or number, just as they had been before, when used in isolation. As stringing words together became a widespread habit, then a norm, it became necessary to disambiguate the subject and object of verbs – very often the speaker or hearer themselves, the most interesting themes for two interacting individuals – with nouns used to address the hearer and self-refer to the speaker. Thanks to this repetitive use, the most frequent of these nouns must, by a process which remains unclear (although probably not forever), have evolved in shortened forms into first- and second-person pronouns.

Our conclusion was that, at the time of Proto-Sapiens, personal pronouns were already being formed, since they are found in all language families (in spite of their not being absolutely necessary, albeit very useful) but were not yet fixed as a lexical category. The original nouns that had given rise to them still varied freely between referring to the speaker and referring to the hearer, according to their original nominal meaning, and only later were fixed onto either first or second person in each family. Since the very existence of first- and second-person pronouns is hardly conceivable without a syntactically articulated language, Proto-Sapiens at the time of its split must have been in the process of acquiring syntactic articulation. This process certainly took time, and perhaps lasted until late into Upper Paleolithic, judging by the fact that there are more reconstructed first- and second-person pronoun roots in ancient taxa, such as Eurasiatic (Greenberg 2000) or Nostratic (Bomhard 2008; Dolgopolsky 2008), than in recent ones. The unexpected absence of a clear-cut distinction between first- and second-person pronominal roots at the global level would thus testify that syntactic articulation had begun to evolve before the dispersion of modern humans, and probably was part of its success, but had not yet led to the development of full-fledged first- and second-person pronouns.

It has been repeated recently that the origin of language is the most difficult scientific problem of our time. At the very least, it is certainly the most difficult problem resisting evolutionary theory. How could one hope to solve it without the powerful tool of comparative linguistics, which opens a window on past times as far back as the initial dispersion of our *Homo sapiens* ancestors? How could one hope to solve it without giving spoken words their legitimate due?

References

- Abaev, V.I. (1970). *Osetinsko-Russkij Slovar' – Iron-Uyryssag Dzyrduat*. Ordžonikidze: IR. <http://ironau.ru/lingvo-abbyy.html> (31 May 2010).
- Academia Română. (1998). *Dicționarul explicativ al limbii române*. Bucharest: Univers Enciclopedic. <http://dexonline.ro/> (15 April 2010).
- Bancel, P.J., & Matthey de l'Etang, A. (2002). Tracing the ancestral kinship system: The global etymon *kaka*. Part I: A linguistic study. *Mother Tongue*, 7, 209–243.
- Bancel, P.J., & Matthey de l'Etang, A. (2005). Kin tongue. A study of kin nursery terms in relation to language acquisition, with a historical and evolutionary perspective. *Mother Tongue*, 9, 171–190.
- Bancel, P.J., & Matthey de l'Etang, A. (2008). The millennial persistence of Indo-European and Eurasiatic pronouns and the origin of nominals. In J.D. Bengtson (Ed.), *In hot pursuit of language in prehistory. Papers in the four fields of anthropology in honor of Harold C. Fleming* (pp. 439–464). Amsterdam: John Benjamins.
- Bancel, P.J., & Matthey de l'Etang, A. (2010). Where do personal pronouns come from? *Вопросы языкового родства / Journal of Language Relationship*, 3, 127–152.
- Bancel, P.J., Matthey de l'Etang, A., & Bengtson, J.D. (2010). Back to Proto-Sapiens (part 2). The global kinship terms *papa*, *mama* and *kaka*. In D. Jones & B. Milicic (Eds.), *Kinship, language, and prehistory. Per Hage and the renaissance in kinship studies* (pp. 38–45). Salt Lake City, UT: University of Utah Press.
- Bengtson, J.D. (Ed.). (2008). *In hot pursuit of language in prehistory. Papers in the four fields of anthropology in honor of Harold C. Fleming*. Amsterdam: John Benjamins.
- Bengtson, J.D., & Ruhlen, M. (1994). Global etymologies. In M. Ruhlen (Ed.), *On the origin of languages. Studies in linguistic taxonomy* (pp. 277–336). Stanford, CA: Stanford University Press.
- Bird, B. (2006). *Online Yaghnobi-Tajik-English lexicon*. <http://yaghnobi.files.wordpress.com/2007/07/yaghnobi-english-tajik-lexicon.pdf> (27 May 2010).
- Boë, L.-J. (2004). *La langue unique et l'hypothèse de Ruhlen. Repères historiques et mise à l'épreuve méthodologique*. Paper presented to the Société de Linguistique de Paris on March 6, 2004.
- Boë, L.-J., Bessière, P., Ladjili, N., & Audibert, N. (2006). Simple combinatorial considerations challenge Ruhlen's mother tongue theory. In B. L. Davis & K. Zajdo (Eds.), *The syllable in speech production* (pp. 63–93). New York, NY: Lawrence Erlbaum Associates.
- Boë, L.-J., Bessière, P., & Vallée, N. (2003). When Ruhlen's "mother tongue" theory meets the null hypothesis. Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona.
- Bomhard, A.R. (2008). *Reconstructing Proto-Nostratic. Comparative phonology, morphology, and vocabulary*. Leiden: Brill.
- Bowler, J.M., Johnston, H., Olley, J.M., Prescott, J.R., Roberts, R.G., Shawcross, W., et al. (2003). New ages for human occupation and climatic change at Lake Mungo, Australia. *Nature*, 421, 837–840.
- Brigaudiot, M., & Danon-Boileau, L. (2002). *La naissance du langage dans les deux premières années*. Paris: Presses universitaires de France.
- Burrow T., & Emeneau, M.B. (1984). *A Dravidian etymological dictionary*. (2nd ed.) Oxford, UK: Clarendon. <http://dsal.uchicago.edu/dictionaries/burrow/> (5 May 2010).
- Buschmann, J.C.E. (1852). Über den Naturlaut. *Philologische und historische Abhandlungen der königlichen Akademie der Wissenschaften in Berlin*, 3, 391–423.

- Cann, R.L., Stoneking, M., & Wilson, A.C. (1987). Mitochondrial DNA and human evolution. *Nature*, 325, 31–36.
- Chantraine, P. (1968). *Dictionnaire étymologique de la langue grecque. Histoire des mots*. Paris: Klincksieck. (<http://www.archive.org/details/Dictionnaire-Etymologique-Grec>) (7 May 2010).
- Charles-Edwards, T.M. (2003). *Early Irish and Welsh kinship*. Oxford, UK: Oxford University Press.
- Ciorănescu, A. (1958–1966). *Dicționarul etimologic român*. Tenerife: Universidad de la Laguna.
- Deacon, H.J. (2001). *Guide to Klasies River*. (<http://academic.sun.ac.za/archaeology/KRguide2001.PDF>) (16 September 2011).
- Deacon, T. (1997). *The symbolic species. The co-evolution of language and the brain*. New York, NY: Norton.
- De Laguna, G. (1927). *Speech: Its function and development*. New York, NY: Yale University Press.
- d'Errico, F., Henshilwood, C.S., Vanhaeren, M., & van Niekerk, K. (2005). *Nassarius kraussianus* shell beads from Blombos Cave: Evidence for symbolic behaviour in the Middle Stone Age. *Journal of Human Evolution*, 48, 3–24.
- de Waal, F. (1982). *Chimpanzee politics. Power and sex among apes*. London: Jonathan Cape.
- Dolgopolsky, A. (1964). Gipoteza drevnejšego rodstva jazykovyx semej severnoj Evrazii s verojatnostnoj točki zrenija, *Voprosy jazykoznanija*, 2, 53–63. [English translation: A probabilistic hypothesis concerning the oldest relationships among language families of Northern Eurasia. In V. V. Shevoroshkin & T. L. Markey (Eds.), *Typology, relationships and time* (pp. 27–50). Ann Arbor, MI: Karoma, 1986].
- Dolgopolsky, A. (2008). *Nostratic dictionary*. McDonald Institute for Archaeological Research, University of Cambridge, UK. (<http://www.dspace.cam.ac.uk/handle/1810/196512>) (9 November 2010).
- Dybo, A. (2006). *Turkic etymology*. Online etymological database, Tower of Babel Project. (<http://starling.rinet.ru/cgi-bin/main.cgi?flags=eygttnl>) (26 May 2010).
- Emeneau, M.B. (1953). Dravidian kinship terms. *Language*, 29, 339–353.
- Fouts, R.S. (1997). *Next of kin. What chimpanzees have taught me about who we are*. New York, NY: William Morrow.
- Greenberg, J.H. (2000–2001). *Indo-European and its closest relatives. The Eurasiatic language family*. Stanford, CA: Stanford University Press.
- Grierson, G. (1920). *Ishkashmi, Zebaki and Yazghulami. An account of three Eranian dialects*. London: Royal Asiatic Society. (<http://www.angelfire.com/sd/tajikistanupdate/engpamirlanguages.html>) (31 May 2010).
- Hayyim, S. (1934–1936). *New Persian-English dictionary*. Tehran: Beroukhim. (<http://dsal.uchicago.edu/dictionaries/hayyim/>) (4 June 2010).
- Henshilwood, C.S., & d'Errico, F. (2005). Being modern in the Middle Stone Age: Individuals and innovation. In C. Gamble & M. Porr (Eds.), *The individual hominid in context: Archaeological investigations of Lower and Middle Palaeolithic landscapes, locales and artefacts* (pp. 244–264). New York, NY: Routledge.
- Henshilwood, C.S., d'Errico, F., Marean, C.W., Milo, R.G., & Yates, R. (2001). An early bone tool industry from the Middle Stone Age at Blombos Cave, South Africa: Implications for the origins of modern human behaviour, symbolism, and language. *Journal of Human Evolution*, 41, 631–678.
- Izard, M. (1965). La terminologie de parenté bretonne. *L'Homme*, 5, 88–100.

- Jakobson, R. (1960). Why *mama* and *papa*? In B. Kaplan & S. Wapner (Eds.), *Perspectives in psychological theory. Essays in honor of Heinz Werner* (pp. 124–134). New York, NY: International Universities Press.
- Kabir, H., & Akbar, W. (1999). *Dictionnaire pashto-français*. Paris: L'Asiathèque.
- Kenneally, C. (2007). *The first word. The search for the origins of language*. New York, NY: Viking.
- Keskin, M. (No date). *Wörterverzeichnis. Zazaki-deutsch deutsch-zazaki*. (http://www.zazaki.de/deutsch/Zazaki-Deutsch_Deutsch-Zazaki_Woerterbuechlein.pdf) (31 May 2010).
- Knight, A., Underhill, P.A., Mortensen, H.M., Zhivotovsky, L.A., Lin, A.A., Henn, B.M., et al. (2003). African Y chromosome and mtDNA divergence provides insight into the history of click languages. *Current Biology* 13–6, 464–473.
- Lieberman, P. (1975). *On the origins of language. An introduction to the evolution of speech*. New York, NY: Macmillan.
- Lieberman, P. (1985). On the evolution of human syntactic ability. Its pre-adaptive bases: Motor control and speech. *Journal of Human Evolution*, 14, 657–668.
- Lieberman, P. (1992). On the evolution of human language. In J.A. Hawkins & M. Gell-Mann (Eds.), *The evolution of human languages* (pp. 21–47). Redwood City, CA: Addison-Wesley.
- Lieberman, P. (2000). *Human language and our reptilian brain. The subcortical bases of speech, syntax, and thought*. Cambridge, MA: Harvard University Press.
- Locke, J.L. (1990). “Mama” and “papa” in child language. Parent reference or phonetic preference? In B. Metuzale-Kangere & H.D. Ringholm (Eds.), *Symposium balticum. A festschrift to honour Professor Velta Ruke-Dravina* (pp. 267–273). Hamburg: Buske.
- Lubbock, J. (1889). *The origin of civilisation and the primitive condition of man. Mental and social condition of savages* (5th ed.). London: Longmans & Green.
- Lubotsky, A. (No date). *Indo-Aryan inherited lexicon*. Online etymological database of the Indo-European Etymological Dictionary project, University of Leiden. (<http://www.indo-european.nl/index2.html>) (20 October 2010) [<http://iedo.brillonline.nl/dictionaries/>].
- MacNeillage, P.F. (2008). *The origin of speech*. Oxford, UK: Oxford University Press.
- MacNeillage, P.F., & Davis, B.L. (1990). Acquisition of speech production: Frame, then content. In M. Jeannerod (Ed.), *Attention and performance*, 13: *Motor representation and control* (pp. 453–476). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Maddieson, I. (1984). *Patterns of sounds*. Cambridge, UK: Cambridge University Press.
- Maddieson, I. (1997). Phonetic universals. In W.J. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences* (pp. 619–639). Oxford: Blackwell.
- Mahadevan, I. (2003). *Early Tamil epigraphy from the earliest times to the 6th century AD*. Cambridge, MA: Harvard University Press.
- Mancino, T. (No date). *Shughni dictionary*. (<http://web.linguist.umass.edu/~woolford/Shughni%20Dictionary.pdf>) (31 May 2010).
- Marean, C.W., Bar-Matthews, M., Bernatchez, J., Fisher, E., Goldberg, P., Herries, A.I.R., et al. (2007). Early human use of marine resources and pigment in South Africa during the Middle Pleistocene. *Nature*, 449, 905–908.
- Matthey de l’Etang, A., & Bancel, P.J. (2002). Tracing the ancestral kinship system: The global etymon *kaka*. Part II: An anthropological study. *Mother Tongue*, 7, 245–258.
- Matthey de l’Etang, A., & Bancel, P.J. (2005). The global distribution of *(p)apa* and *(t)ata* and their original meaning. *Mother Tongue*, 9, 133–169.
- Matthey de l’Etang, A., & Bancel, P.J. (2008). The age of *mama* and *papa*. In J.D. Bengtson (Ed.), *In hot pursuit of language in prehistory. Papers in the four fields of anthropology in honor of Harold C. Fleming* (pp. 417–438). Amsterdam: John Benjamins.

- Matthey de l'Etang, A., & Bancel, P.J. (2012). The Proto-Amerind kinship system. Communication presented at the American Anthropological Association 111th Annual Meeting, held in San Francisco, CA, November 14–18, 2012.
- Matthey de l'Etang, A., Bancel, P.J., & Ruhlen, M. (2010). Back to Proto-Sapiens (Part 1). The inherited kinship terms *papa*, *mama* and *kaka*. In D. Jones & B. Milicic (Eds.), *Kinship, language, and prehistory. Per Hage and the renaissance in kinship studies* (pp. 29–37). Salt Lake City, UT: University of Utah Press.
- McDougall, I., Brown, F.H., & Fleagle, J.G. (2005). Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature*, 433, 733–736.
- Meyer, G. (1891). *Etymologisches Wörterbuch der albanesischer Sprache*. Strasbourg: Trübner. <<http://www.archive.org/stream/etymologischesw00meyegoog/>> (16 June 2010).
- Meyer-Lübke, W. (1911). *Romanisches etymologisches Wörterbuch*. Heidelberg: Carl Winter. <<http://ia311218.us.archive.org/2/items/romanischesetymo00meyeuoft/romanischesetymo00meyeuoft.pdf>> (15 April 2010).
- Mumtaz, A. (1985). *Baluchi glossary. A Baluchi-English glossary. Elementary level*. Kensington, MD: Dunwoody. <<http://dsal.uchicago.edu/dictionaries/mumtaz/>> (31 May 2010).
- Murdock, G.P. (1959). Cross-language parallels in parental kin terms. *Anthropological Linguistics*, 1, 1–5.
- Nebrija, E.A. de. (1492). *Lexico o diccionario latino-español*. Online facsimile. <<http://www.cervantesvirtual.com/servlet/SirveObras/07032774389636239647857/ima0192.htm>> (12 June 2010).
- Nikolayev, S. (2007). *Indo-European etymology*. Online etymological database, Tower of Babel Project. <<http://starling.rinet.ru/cgi-bin/main.cgi?flags=eygtnnl>> (26 May 2010).
- Nişanyan, S. (2001). *Sözlerin Soyağacı. Çağdaş Türkçenin Etimolojik Sözlüğü*. Istanbul: Everest. <<http://www.nisanyansozluk.com/>> (12 June 2010).
- O'Connell, J.F., & Allen, J. (2004). Dating the colonization of Sahul (Pleistocene Australia-New Guinea): A review of recent research. *Journal of Archaeological Science* 31, 835–853.
- Oller, D.K. (1980). The emergence of the sounds of speech in infancy. In G.H.Yeni-Komshian, J.F. Kavanagh, & C.A. Ferguson (Eds.), *Child phonology: Vol. 1, Production* (pp. 93–112). New York, NY: Academic Press.
- Pagel, M. (2000). Maximum likelihood models for glottochronology and for reconstructing linguistic phylogenies. In C. Renfrew, A. McMahon, & L. Trask (Eds.), *Time depth in historical linguistics* (pp. 189–207). Cambridge, UK: McDonald Institute for Archaeological Research.
- Pehrson, R.N. (1966). *The social organization of the Marri Baluch*. New York, NY: Wenner-Gren Foundation for Anthropological Research.
- Pokorný, J. (1959). *Indogermanisches etymologisches Wörterbuch*. Bern: Francke.
- Raverty, H.G. (1867). *A dictionary of the Puk'hto, Pus'hto, or language of the Afghans. With remarks on the originality of the language, and its affinity to other oriental tongues* (2nd ed.). London: Williams and Norgate. <<http://dsal.uchicago.edu/dictionaries/raverty/>> (31 May 2010).
- Ringe, D.A. (2002). Review of Greenberg (2000). *Journal of Linguistics*, 38, 415–420.
- Rosenfelder, M. No date. Numbers in 5020 Languages. In *The Metaverse*. <<http://www.zompist.com/numbers.htm>> (21 March 2011).
- Ruhlen, M. (1994a). *The origin of language. Tracing the evolution of the mother tongue*. New York, NY: John Wiley.
- Ruhlen, M. (1994b). *On the origin of languages. Studies in linguistic taxonomy*. Stanford, CA: Stanford University Press.

- Rybatzki, V. (2006). *Die Personennamen und Titel der mittelmongolischen Dokumente. Eine lexikalische Untersuchung*. Ph.D. dissertation. Helsinki: University of Helsinki. <https://oa.doria.fi/bitstream/handle/10024/828/dieperso.pdf?sequence=1> (26 May 2010).
- Schulze, W. (2000). *Northern Talysh*. Munich: Lincom.
- Schurmann, F. (1962). *The Mongols of Afghanistan. An ethnography of the Moghòls and related peoples of Afghanistan*. The Hague: Mouton.
- Singer, R., & Wymer, J. (1982). *The Middle Stone Age at Klasies River Mouth in South Africa*. Chicago, IL: University of Chicago Press.
- Sköld, H. (1936). *Materialien zu den iranischen Pamirsprachen*. Lund: Gleerup.
- Starostin, S. (2006). *Chinese characters*. Online etymological database, Tower of Babel Project. <http://starling.rinet.ru/cgi-bin/main.cgi?flags=eygtnnl> (26 May 2010).
- Stern, C., & Stern, W. (1907). *Die Kindersprache: Eine psychologische und sprachtheoretische Untersuchung*. Leipzig: Barth.
- Strand, R. (1997–2012). Kinship systems of the Hindu-Kush. In R. Strand, *Nuristân. Hidden Land of the Hindu-Kush*. <http://nuristan.info/> (3 May 2010).
- Trask, L. (2004). Where do *mama/papa* words come from? *University of Sussex Working Papers in Linguistics and English Language*. <http://www.sussex.ac.uk/english/research/projects/linguisticspapers> (20 September 2013).
- Turner, R.L. (1962–1966). *A comparative dictionary of Indo-Aryan Languages*. London: Oxford University Press. <http://dsal.uchicago.edu/dictionaries/soas/> (2 May 2010).
- Vocabularium Cornicum*. 12th century anonymous manuscript. <http://www.carlaz.com/cornish/voccorn.txt> (12 April 2010).
- Wang, W. S.-Y. (2004). *Chinese dialects*. Online etymological database, Tower of Babel Project. <http://starling.rinet.ru/cgi-bin/main.cgi?flags=eygtnnl> (16 May 2010).
- Werner, B. (2009). *Terminologie der erweiterten Familienstruktur der Zaza (Region Çermik-Gerger-Siverek)*. http://www.zazaki.de/zazaki/Kinshipdiagram_deutsch.pdf (31 May 2010).
- Westermarck, E. (1891). *The history of human marriage*. London: Macmillan.
- White, T.D., Asfaw, B., DeGusta, D., Gilbert, H., Richards, G.D., Suwa, G., et al. (2003). Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature*, 423, 742–747.

Appendices: Comparative data

Appendix A. The Proto-Indo-European root **ma-* ~ **mama-* ‘mother’ [or, rather, ‘mother, mom’]

The reference of data not drawn from Nikolayev (2007) is given when it is relatively difficult to access (i.e. essentially for the Indic, Nuristani, and Iranian groups); additional data from European language groups have been drawn from standard dictionaries, often accessible on the Internet.

Indic: Proto-Indic **mā* ‘mother’: Pali *māmikā* ‘mother’; Prakrit *māu* ‘mother’; Germany Gypsy *mama* ‘mother’; Romania Gypsy *māmi* ‘grandmother’; Bashkarik *mēm* ‘mother’s mother’, *mām* ‘mother’s father’; Phalula *mēmi* ‘mother’s mother’, *mómo* ‘mother’s father’; Domaki *māma* ‘mother’; Tirahi *mā* ‘mother’; Shina (Gilgiti dial.) *mā* ‘mother’; Shina (Kohistani, Palesi) *mā* ‘mother’; Shina (Guresi) *māh*

'mother'; Sindhi *māu* 'mother'; Lahnda *mā* 'mother'; Lahnda (Awankari dial.) *mā* 'mother'; Punjabi *mā̃ ~ māū ~ māi ~ māmmī* 'mother'; West Pahari (Curahi dial.) *mā* 'mother'; Kotgarhi *mā* 'mother', *māi* 'mother, goddess Durga'; Kumauni *mā* 'mother, mother-in-law'; Nepali *māu* 'female animal having given birth'; Assamese *mā ~ māu* 'mother', *māi* 'mother, mother's brother's wife'; Bengali *mā* 'mother', *māi* 'breast'; Oriya *māā ~ mā* 'mother', *māi* 'woman'; Maithili *māi* 'mother'; Bhojpuri *māi* 'mother'; Awadi (Lakhimpuri dial.) *māi* 'mother'; Hindi *mā ~ māi ~ mā̃* 'mother'; Old Marwari *mā* 'mother'; Gujarati *mā ~ māi* 'mother'; Marathi *mā ~ māi* 'mother', *māi* 'mother-in-law' (Turner 1962–1966, etym. 10016 & 10058).

Iranian: Ossetic *mamæ* 'mom' (Abaev 1970); Yaghnobi *momo* 'grandmother' (Bird 2006); Wakhi *mum* 'grandmother' (Grierson 1920); Persian *mām* 'mom', *māmā* 'midwife' (Hayyim 1934–1936); Zaza *ma* 'mother' (Werner 2009).

Armenian: *mam* 'grandmother'.

Hellenic: Classical Greek *mā gā* 'Earth Mother!'; (Homeric) *māia* 'address to an old woman'; (Attic) *māia* 'mom, wet nurse, midwife'; *māmmē* 'mom, granny'; (Doric) *māia* 'granny'; Standard Modern Greek *mama* 'mom', *mammi* 'granny'.

Slavic: Proto-Slavic **mama* 'mom': Belorussian *mama* 'mom'; Russian *mama* 'mom'; Ukrainian *mama* 'mom'; Bulgarian *mama* 'mom'; Serbo-Croatian *mama* 'mom'; Slovene *mama* 'mom'; Czech *mama* 'mom'; Slovak *mama* 'mom'; Polish *mama* 'mom'; Upper Sorbian *mama* 'mom'; Lower Sorbian *mama* 'mom'.

Baltic: Proto-Baltic **mama* 'mom'; Lithuanian *mama*, (dial.) *mōmā* 'mom'; Latvian *māma* 'mom'.

Germanic: Standard German *Mama* 'mom', *Oma* 'granny'; Alemannic *Mamme* 'mom'; Alsatian *Mamma* 'mom'; Low German *Mame ~ Mamme ~ Mammā* 'mom'; Dutch *ma ~ mam ~ mama* 'mom', *oma* 'granny'; Danish *mama* 'mom'; Swedish *mamma* 'mom'; Norwegian *mamma* 'mom'; Faeroese *mamma* 'mom'; Icelandic *mamma* 'mom'.

Italic: Latin *Maia* 'Great Goddess = Earth, associated with the cult of Vulcan, and mother of Mercury', *Maius* 'month of May', *mamma* 'mommy, mother, wet nurse'; Romanian *mămă* 'mother, mom'; Italian *mamma* 'mom'; Sursilvan *mumma* 'mom'; Sutsilvan *moma* 'mom'; Surmiran *mamma* 'mom'; Putér *mamma* 'mom'; Vallader *mamma* 'mom'; Friulian *mame* 'mom'; French *maman* 'mom', *mémé ~ mamie* 'granny'; Occitan *mamà* 'mom', *mameta* 'granny'; Catalan *mamà* 'mom'; Spanish *mamá* 'mom'; Portuguese *mamãe* 'mom'.

Celtic: Proto-Celtic **mamā*: Old Irish *mam* 'mother'; Welsh *mam* 'mother'; Breton *mam* 'mother'; Cornish *mam* 'mother'; Proto-Celtic **mammiā*: Old Irish *muimme* 'foster mother'.

Albanian: Tosk *memë* 'mother'; Gheg *mame* 'mother'.

Appendix B. The Proto-Indo-European root **pa ~ *papa* 'father, dad'

Anatolian: Palaic *pāpa* 'father'.

Indic: 1. Proto-Indic **bappa*: Prakrit *bappa* 'father'; Armenia Gypsy *bap* 'father'; Dameli *bāp* 'father, grandfather'; Gawar-Bati *bāp* 'father'; Torwali *bāp* 'father'; Lahnda *bāpū* 'grandfather'; Punjabi *bāp, bāpū* 'father'; Nepali *bāp* 'father'; Assamese *bāp* 'father'; *bāpā* 'term of address to a father or of affection to a young man', *bāpu* 'term of address to a learned Brahman'; Bengali *bāp* 'father', *bāpu* 'father, child'; Oriya *bāpa* 'father', *bapā* 'term of endearment to younger persons', *bāpu* 'term of address

to a father or to a young person, (Puri dial.) *bāpā* ‘father’s father’; Maithili *bāp*, *bappā* ‘father’; Awadi (Lakhimpuri dial.) *bāp* ‘father’; Hindi *bāp* ‘father’; Gujarati *bāp* ‘father’; Marathi *bāp* ‘father’; Sinhalese *bapa* ‘father’; West Pahari (Koci dial.) *bāp* ‘father’, (Kiuthali) *bapu* (used by Rajputs), *bāpū*; Maldivian (upper class) *bappa*, (lower class) *bafā* ‘father’. 2. Proto-Indic **babba*: Domaki *baba* ‘father, father’s brother’ (pl. *piāra* < *pitṛ*); Pashai (Areti dial.) *bāba* ‘father’; Shumashti *bābā*; Bashkarik *bab* ‘father’, *bobā* ‘father’s brother’; Savi *bāb*, *bābu* ‘father’; Phalula *bābu* ‘father’, *bābā* ‘father’s brother’; Shina (Gilgiti dial.) *bābu* ‘father’, (Palesi) *bubā*; Kashmiri *bab* ‘father, grand-father’, *bāb* ‘father’, (Rambani dial.) *babb* ‘father’, (Poguli) *baub* ‘father’, (Dodi) *babbō* ‘father’; Punjabi *bābbā* ‘father, grandfather’, *bābū* ‘term of respect’, (Kangra dial.) *babb* ‘father’; West Pahari (Bhadrawahi dial.) *bābō* ‘father’, (Bhalesi) *bāb* ‘father’, (Curahi) *bābb* ‘father’, (Cameali) *babb* ‘father’, (Khashali) *babb* ‘father’ (voc. *bāvā*); Kumauni *bābu* ‘father’, *babā* ‘affectionate term for father or child’; Nepali *bābu* ‘father’, *bābai* ‘term of address to child’, *babuwā* ‘father’, (Tarai dial.) affectionate term for son’; Bengali *bābā* ‘father, baby’, *bābu* ‘gentleman’; Oriya *bābā* ‘father’, *babā* ‘father’s elder brother’, *bābu* ‘gentleman’, *babuā* ‘term of endearment to juniors’; Maithili *bābā* ‘father’, *bābu* ‘title of respect’; Hindi *bābu* ‘father’, *babuwā* ‘child’; Gujarati *bābū* ‘term of respect’; Marathi *bābū* ‘term of respect’; Marathi *bābdā* ‘term of endearment to a child’; West Pahari (Koci dial.) *bāb* ‘father’, (Kiuthali) *babu* ‘father’ (used by Rajputs), *bābū* ‘father’. (Turner 1962–1966: etym. 9209)

Nuristani: Kāta-vari (Ktivi dial.) *vov* ‘grandfather’; Kām’iri *vov* ‘grandfather’; Supu-vari *vā* ‘grandfather’; Sanu-viri *bāba* ‘elder brother’; Usüt-vare *vāv* ‘grandfather’, *bab* ‘elder brother’; Vā-alā *bāba* ‘elder brother’; Ameš-alā *bāba* ‘elder brother’; Nišei-alā *bāba* ‘elder brother’. (Strand 1997–2008)

Iranian: Khwarezmian *papa*, *bāb* ‘father’ (Rybatzki 2006); Sogdian *bābay* ‘father’ (Rybatzki 2006); Yaghnobi *bobō* ‘grandfather’ (Bird 2006); Bactrian *babu* ‘masc. personal name’ (Rybatzki 2006); Pashto *bābū* ‘dad, address term to an elder’, *bābā* ‘grandfather’ (Kabir & Akbar 1999; Schurmann 1962); Wakhi *pūp* ‘grandfather’ (Grierson 1920); Sanglechi *bobō* ‘father’s father’ (Rybatzki 2006); Ishkashmi *bōbō* ‘grandfather’ (Grierson 1920); Shughni *būb* ‘grandfather’ (Sköld 1936); Bajui *bōb* ‘grandfather’ (Sköld 1936); Sahdara *bōb* ‘grandfather’ (Sköld 1936); Bartangi *bōb* ‘grandfather’ (Sköld 1936); Yazghulami *bāb* ‘grandfather’ (Sköld 1936); Parachi *bāw* ‘father’, *bābā* ‘grandfather’ (Rybatzki 2006); Pahlavi *bābā* ‘first part of masc. name’ (Rybatzki 2006); Farsi *bābā* ‘father, grandfather’ (Rybatzki 2006); Basseri Farsi *ba*⁰ ‘father’, *bābō* ‘grandfather’ (Rybatzki 2006); Dari *bābā* ‘grandfather, father, dad’ (Rybatzki 2006); Tajik *baba* ~ *bawa* ~ *baab* ‘father’, *bābā* ‘ancestor’ (Schurmann 1962); Baluchi *bābā* ‘elder man’ (Rybatzki 2006); Marri Baluch *baba* ‘father, grandfather, grandfather’s brother’ (Pehrson 1966); Hazara *bābā* ‘father’ (Schurmann 1962); Kurdish *bav* ‘father’, *bavo* ‘dad’, *bapir* ‘grandfather’ (Rybatzki 2006); Zaza *baō* ‘dad (vocat.)’ (Keskin no date).

Armenian: *pap* ‘grandad’.

Hellenic: Classical Greek *pappa* ‘dad’, *pappos* ‘grandfather, forebear, ancestor’; Modern Pontic Greek *papa* ‘dad’ (Fauvin & Nikaki, personal communication); Standard Modern Greek *baba* ‘dad’ (borrowed from Turkish, see Chantraine 1968), *pappoús* ‘grandfather’.

Baltic: Latvian *paps* ‘dad’.

Germanic: Gothic *papa* 'dad'; Modern High German *Papa* 'dad', *Opa* 'grandad'; Alsatian *Pàpa* 'dad'; Alemannic *Pappe* 'dad'; Rhine Franconian *Pàppe* ~ *Bàbbe* 'dad'; Bavarian *Babba* 'dad'; Dutch *pa* ~ *papa* ~ *pappa* 'dad', *opa* 'grandad'; English *papa*; Danish *papa* 'dad'; Swedish *pappa* 'dad'; Norwegian *pappa* 'dad'; Faeroese *pápi* 'dad'; Icelandic *pabbi* 'dad'.

Italic: Latin *pappa* 'dad', *pappus* 'grandfather, ancestor'; French *papa* 'dad', *pépé* ~ *papy* 'grandad'; Sursilvan *bab* 'father'; Sutsilvan *bab* 'father'; Surmiran *bab* 'father'; Putér *bap* 'father'; Vallader *bap* 'father'; Friulian *pai* 'dad'; Italian *papà* ~ *babbo* 'dad'; Occitan *papà* 'dad', *papet* 'grandad'; Catalan *papà* 'dad'; Spanish *papá* 'dad'; Portuguese *pai* ~ *papá* ~ *papai* 'dad'.

Albanian: *baba* 'dad' (borrowed from Turkish, Meyer 1891).

Appendix C. The Proto-Indo-European root **tat-* ~ **tet-* 'father' [or, rather, **tata* 'dad, father']

Anatolian: Hieroglyphic Luwian *tati(a)-* 'father'; Luwian *tati(ja)-* 'father'; Lycian *tedi* 'father'.

Indic: Sanskrit *tātá* '(vocative) affectionate address to junior' (Śatapatha Brāhmaṇa), 'idem to senior' (Mahābhārata), 'father' (ibid.), *tatá* 'father' (Rig Veda); Pali *tāta* 'term of respectful or affectionate address to an elder or younger'; Prakrit *tāa* 'father, son'; Germany Gypsy *tatta* 'father'; Pašai (Darrai-i Nūr and Wegali dial.) *tati* 'father'; Khowar *tat*; Old Gujarati *tāya* m. (Turner 1962–1966: etym. 5754). *Proto-Indic* **dādda* 'father or other elderly relative': Germany Gypsy *dād* 'father'; Domaki *dādo* 'grandfather'; Dameli *dādi* 'father'; Pašai (Laurowani dial.) *dādā* 'elder brother', (Gulbahari) *dādā* 'father', (Kurangali) *dādo* 'father's brother'; Kalasha *dāda* 'father'; Bashkarik *dād* 'grandfather', *dēd* 'grandmother'; Phalula *dōdo* 'father's father', *dēdi* 'father's mother'; Shina *dādo* 'grandfather', *dādi* 'grandmother'; Sindhi *ḍāḍō* 'father's father', *ḍāḍi* 'father's mother', (Kacchi dial.) *ḍāḍo* 'grandfather'; Lahnda *ḍāḍā* 'father's father', *ḍi* 'father's mother', *dādā* m., *dī* f.; Punjabi *dāddā*, *dā* m., *dāddi*, *dī* f.; Western Pahari (Bhalesi dial.) *dādo* m., (Kotgarhi) *dād* 'father's father, elder brother', *daddi* 'father's mother', (Kiuthali) *dādā* 'grandfather'; Kumauni *dādā* 'grandfather, elder brother', *dādi* 'grandmother, elder sister', *dā* 'address to an elder brother'; Nepali *dādā* 'old servant', *dājyu*, *dāi* (contaminated by *bhāi* < *bhrātṛ* [Proto-Indic form of PIE **bhratēr* 'brother', PJB & AME]) 'elder brother'; Assamese *dādā* 'elder brother'; Bengali *dādā* 'grandfather, elder brother', *dādi* 'grandmother'; Oriya *dādā* 'grandfather, father's brother, elder brother'; Maithili *dādā* 'grandfather'; Hindi *dādā* 'father's father, elder brother', *dādi* 'father's mother'; Gujarati *dādō* 'father's father', *dādi* f.; Marathi *dādā* 'elder brother', *dādi* 'respectful term for an old woman' (Turner 1962–1966: etym. 6261).

Nuristani: Kāta-vari (Ktivi dial.) *to* 'father', *-to* 'father's (brother)'; Kāmv'iri *tot* 'father', *-tot* 'father's (brother)'; Vā-alā *tāta* 'father', *-ta* 'father's (brother)', *el-ta* 'grandfather' (cf. *ei* 'mother', *el-ei* 'grandmother'); Ameš-alā *tāta* 'father', *-tāta* 'father's (brother)', *gaṇ-ta* 'grandfather' (cf. *gaṇ-ei* 'grandmother'); Niše-alā *tāti* 'father', *-tāti* 'father's (brother)' (Strand 1997–2008).

Iranian: Old Avestan *tā* 'father' (Yasna 47.3); Yaghnobi *dodo* 'father' (Bird 2006); Shughni *tat* 'father' (Mancino no date); Roshani *taat* 'father', *tatek* 'grandfather'; Ishkashmi *tot* ~ *tāt* 'father' (Grierson 1920); Wakhi *tat* 'father' (Grierson 1920); Zebaki

tāt ~ *tā* ‘father’ (Grierson 1920), Pashto *dādā* ‘a term of endearment to a father or elder brother (East), also elder sister (West)’ (Raverty 1867); Zaza *ded* ‘father’s brother’, *dedo* ‘idem (voc.)’ (Werner 2009); Talysh *dada* ‘father’ (Schulze 2000); Baluchi *dada* ‘father’s father’ (Mumtaz 1985).

Greek: Classical Greek (Myrin.) *tatā* (voc.) ‘daddy’, Homeric *tétta*.

Slavic: Proto-Slavic **tata* ‘father, dad’: Pskov, Arkhangelsk, Eastern and Southern dialects of Russian *tata* ‘father, dad’; Bulgarian *tato* ‘father, dad’; Serbo-Croatian *táta* ‘father, dad’; Slovenian *táta* ‘father, dad’; Czech *táta* ‘father, dad’; Polish *táta* ‘father, dad’; Lower Sorbian *táta* ‘father, dad’; Upper Sorbian *táta* ‘father, dad’.

Baltic: Proto-Baltic **tet-ia-*, **tēt-iā-*: Lithuanian *tėti-s* ‘father’, *tētē* ‘dad’; East Lithuanian *tētē* ‘father’; Samogitian *titi-s*, dial. *táiti-s* ‘father’; Latvian *tēte*, *tētis* ‘dad’.

Italic: Latin *tāta* ‘dad’; Old Castilian *taita* ‘dad’ (Nebrija 1492), Old Catalan *taita* ‘dad’; Catalan (dialectal) *tata* ‘dad, brother’; Neapolitan *tata* ‘dad’; Romanian *tátă* ‘father, dad’; Sursilvan *tat* ‘grandfather’; Sutsilvan *tat* ‘grandfather’; Surmiran *tat* ‘grandfather’.

Celtic: Old Cornish *tat* ‘father’ (*Vocabularium Cornicum* c. 1250); Cornish *tat* ‘father’; Middle Welsh *tad* ‘father’ (Charles-Edwards 2003); Welsh *tad* ‘father’, *dada* ‘dad’; Middle Breton *tat* ‘father’ (Izard 1965); Breton *tad* ‘father’, *tata* ‘dad’; Old Irish *data* ‘foster father’ (Charles-Edwards 2003).

Albanian: *tate* ‘father’.

Appendix D. The Proto-Dravidian root **appa* ‘dad, father’

Tamil *appan*, *appu* ‘father (term of endearment used to little children or inferiors)’, *appacci* ‘father’, *appāttai* ‘elder sister’, *appi* ‘mistress of house, elder sister’; Malayalam *appan* ‘father’, *appu* ‘affectionate appellation of boys’; Kannada *appa* ‘father (frequently added to the proper names of men as a term of common respect; used endearingly to children by elders)’, *apa* ‘father’, *appu* ‘affectionate appellation of boys’; Kodagu *appē* ‘father’; Tulu *appa*, *appē* ‘affix of respect added to proper names of men’, *appē* ‘mother’, *appa* ‘a mode of calling a mother’; Telugu *appa* ‘father, mother, elder sister (frequently added to names of men as a term of common respect)’; Kolami *appa* ‘father’s sister’; Gondi *āpōrāl* ‘father’, *maipō* ‘my father’, *mī-āpō* ‘thy father’; Maria *tape* ‘father’; Konda *tappe*, (L.) *tāpe* ‘father’ (Voc. 1656); Koya Su. *tappe* ‘(his, her) father’; Konda *aposi* ‘father (with reference to third person)’ (Burrow & Emeneau 1984, etym. 156).

Appendix E. The Proto-Turkic roots **ata* ‘dad, father’, **apa* ‘dad, father’, and **ana* ‘mom, mother’

1. Proto-Turkic **ata* ‘father’: Old Uighur *ata* ‘father’; Sary-Yughur *ata* ‘father’; Nogai *ata* ‘father’; Oirat *ada* ‘father, ancestor’; Karakhanid *ata* ‘father’; Turkmen *ata* ‘father’s father’; Azeri *ata* ‘father’; Balkar *ata* ‘father’; Tuvin *a'da* ‘father’; Middle Turkish *ata* ‘father’; Tatar *ata* ‘father’; Kumyk *ata* ‘father’; Tofalar *ada* ‘father’; Uzbek *ota* ‘father’; Kirghiz *ata* ‘father, ancestor’; Karakalpak *ata* ‘ancestor’; Modern Turkish *ata* ‘ancestor’; Bashkir *ata* ‘father’; Uighur *ata* ‘father, ancestor’; Urum *ata* ‘father’; Cuman *ata* ~ *atta* ‘father’; Kazakh *ata* ‘father’; Khakassian *ada* ‘father’; Karaïm *ata* ‘ancestor’ (Dybo 2006).

2. *Proto-Turkic *apa* 'father': Orkhon *apa* 'ancestors'; Old Uighur *apa* 'ancestors'; Salar *aba* ~ *apa* 'father'; Bashkir (dial.) *apa* 'father'; Sary-Yughur *awa* 'father'; Khakassian *aba* 'father'; Karakhanid *apa* 'father, ancestor'; Tatar (dial.) *aba* 'father'; Tuvin *ava* 'father'; Turkish *aba* 'father'; Kirghiz *aba* 'father'; Altai *aba* 'father, bear'; Azeri (dial.) *aba* 'father'; Balkar *appa* ~ *aba* 'father'; Chuvash *oba* 'bear'; Turkmen (dial.) *aba* 'father'. (Dybo 2006)
3. *Proto-Turkic *ana* 'mother, mom': Old Uighur *ana* 'mother'; Karakhanid *ana* 'mother'; Azeri *ana* 'mother'; Dolgan *iñe* 'mother'; Gagauz *ana* 'mother'; Turkmen *ana* 'mother'; Tuvin *ije* 'mother'; Karaim *ana* 'mother'; Middle Turkish *ana* 'mother'; Khakassian *inä* 'mother'; Kirghiz *ene* 'mother'; Karakalpak *ana* 'mother'; Oirat *ene* 'mother'; Kazakh *ana* 'mother'; Salar *ana* 'mother'; Uighur *ana* 'mother'; Chuvash *añne* 'mother'; Bashkir *inä* 'mother'; Sary-Yughur *ana* 'mother'; Kumyk *ana* 'mother'; Yakut *ije* 'mother'; Balkar *ana* 'mother'. (Dybo 2006)

Appendix F. The origin of words for 'dad', 'father', 'mom', and 'mother' in the Chinese family.

The two databases (Starostin 2006; Wang 2004) from which the following data have been drawn differ in their respective transcriptions of oral stops; we have aligned them according to Wang's transcription. Superscript numbers following modern dialectal forms in Wang's data transcribe tonal contours.

1. 'Dad': Preclassic *paʔ* 'father'; Classic *pá*; Western Han *pá*; Eastern Han *pwá*; Early Postclassic *pwó*; Middle Postclassic *pwó*; Late Postclassic *pwó*; Middle Chinese *pwá* (Starostin 2006). *Modern forms*: Beijing *pa³*; Jinan *pa³*; Xi'an *pa³*; Taiyuan *pa³*; Hankou *pa¹²*; Chengdu *pa¹²*; Yangzhou *pai³*; Suzhou *pa¹¹*; Wenzhou *pa¹¹*; Changsha *pa⁴*; Shuangfeng *po¹¹*; Nanchang *pak⁴¹*; Meixian *pa¹¹*; Guangzhou *pa¹¹*; Xiamen *pa³²*; Chaozhou *pa¹¹*; Fuzhou *pa¹¹*; Shanghai *pa¹* (Wang 2004).
2. 'Father': Preclassic *pahʔ*; Classic *pá*; Western Han *pá*; Eastern Han *pwá*; Early Postclassic *pwó*; Middle Postclassic *pwó*; Late Postclassic *pwó*; Middle Chinese *pú* (Starostin 2006). *Modern forms*: Beijing *fu³*; Jinan *fu³*; Xi'an *fu³*; Taiyuan *fu³*; Hankou *fu³*; Chengdu *fu³*; Yangzhou *fu³*; Suzhou *fu³²*; Wenzhou *voy²²*; Changsha *fu³¹*; Shuangfeng *ɣu³²*; Nanchang *ɬu³²*; Meixian *fu³*; Guangzhou *fu³²*; Xiamen *hu³²* (lit.), *pe³²*; Chaozhou *pe²²*; Fuzhou *xu³²*; Shanghai *vu³²*; Zhongyuan yinyun *fu²* (Wang 2004).
3. 'Mom': Preclassic *m^hāʔ*; Classic *m^hā*; Western Han *m^hā*; Eastern Han *m^hā*; Early Postclassic *m^hō*; Middle Postclassic *m^hō*; Late Postclassic *m^hō*; Middle Chinese *mó* (Starostin 2006). *Modern forms*: Beijing *ma¹¹*; Jinan *ma¹¹*; Xi'an *ma¹¹*; Taiyuan *ma¹*; Hankou *ma¹¹*; Chengdu *ma¹¹*; Yangzhou *ma¹¹*; Suzhou *ma¹¹*; Wenzhou *ma²²*; Changsha *ma¹¹*; Shuangfeng *mo¹¹*; Nanchang *ma¹¹*; Meixian *ma¹¹*; Guangzhou *ma¹¹*; Xiamen *ma¹¹*; Chaozhou *ma¹¹*; Fuzhou *ma¹¹*; Shanghai *ma¹*; Zhongyuan yinyun *ma²* (Wang 2004).
4. 'Mother': Preclassic *māʔ*; Classic *mǎ*; Western Han *mǎ*; Eastern Han *mǎ*; Early Postclassic *mó*; Middle Postclassic *mǎw*; Late Postclassic *mǎw*; Middle Chinese *māw* 'mother' (Starostin 2006). *Modern forms*: Beijing *mu²*; Jinan *mu²*; Xi'an *mu²*; Taiyuan *mu²*; Hankou *mu²*; Chengdu *mu²*; Yangzhou *mo²*; Suzhou *mo¹¹*; Wenzhou *mo²²*; Changsha *mo²*; Shuangfeng *mu²*; Nanchang *mu²*; Meixian *mu¹¹*; Guangzhou *mou²²*; Xiamen *bu²* (lit.), *bo²*; Chaozhou *bo²¹*; Fuzhou *mu²*; Shanghai *mu³*; Zhongyuan yinyun *mu²* (Wang 2004).

Appendix G. The descent of Proto-Indo-European **deikē*- ‘to show, to point’ and **dekṃ*- ‘ten’.

1. Proto-Indo-European **deikē*- ‘to show, to point’:

Anatolian: Hittite: *tekk-ussai*- ‘to show’.

Indic: Proto-Indic **diś* ‘to point’:

(a) Sanskrit (Rig Veda) *diś-āti* ‘points out’, *diś* ‘direction, region’, (Mahābhārata) *diś-ā* ‘direction’; Pali *diś-āti* ‘points out’, *disā* ‘id.’, (Vājasaneyi-Samhitā) *dēśā* ‘point, region, part’; (Rāmāyaṇa) *dēśā* ‘province, country’; Prakrit *dis-ai* ‘tells’, *dis-ā* ‘direction’, *dēsa* ‘part, country’; Old Gujarati *dis-i* ‘direction’; Old Awadhi *dih* ‘direction’, *desa* ‘country’; Armenia Romany *les* ‘earth, world, life’; Palestine Romany *dēs* ‘place, camping ground’; Kalasha (Rumbur dial.) *dēš* ‘country’, *dēša* ‘far, distant’; Phalula *dēš* ~ *diš* ‘village’; Pashai *dēš* ‘cultivated field’; Torwali *diš-ā* ‘towards’; Kashmiri (Kashtawari dial.) *diś* ‘country’; Shina *diš* ‘place’; Sindhi *dehu* ‘country’; Western Pahari (Bhadrawahi dial.) *dēš* ‘village’; Kumauni *des* ‘country’; Nepali *des* ‘country, plains of India’; Bengali *des* ‘country’; Oriya *desa* ‘country’; Maithili *des* ‘country’; Assamese *dih* ‘means, direction’; Hindi *dis-na* ‘to show, to exhibit’, *dis* ‘direction, side’, *des* ‘country’; Marwari *des-ro* ‘small country’; Gujarati *des* ‘country’; Marathi *des* ‘country’; Sinhalese *das-aya* ‘direction’, *desa* ‘country’. (Turner 1962–1966: etym. 6339, 6340, 6547)

(b) Sanskrit (Kauśikasūtra) *diṣ-ṭi* ‘a measure of length’; Shina (Gilgiti dial.) *di-ṭ* ‘span’, (Jijelut dial.) *diṣ* ‘span’; Dameli *diṣ-ṭ* ‘span’; Khowar *diṣ-ṭ* ‘handspan’; Kalasha *jiṣ-ṭ* ‘handspan’; Phalula *diṣ-ṭ* ‘span’. (Turner 1962–1966: etym. 6343)

Nuristani: Ashkun *deši* ‘village’; Kalasha-ala (Waigali) *deš* ‘village’. (Turner 1962–1966: etym. 6340)

Iranian: (a) Old Avestan *ā-diš-ti*- ‘direction’; Avestan *daēs* ‘to show’, *dax-šta*- ‘sign, revelation’; Khotanese *dis*- ‘to confess’; Sogdian *p-ḏ’ys* ‘to show’; Parthian *dyš-g* ‘sign’; *b-dys*- ‘to show’; Ossetic *dis* ~ *des* ‘amazement’; *æv-dis-yn* ~ *æv-des-un* ‘to show’. (Lubotsky no date).

(b) Avestan *diš-ti* ‘a measure of length’; Khotanese *di-ṭhi* ‘a measure of length’.

Greek: (a) Class. Greek *deik-numi* ‘to show’; Cretan *dik-nuti* ‘to show’; Mod. Greek *deix-no* ‘to show’.

(b) Classical Greek (?) *dak-tulos* ‘finger’; Modern Greek (?) *ḍak-tilo* ‘finger’.

Germanic: Proto-Germanic (a) **ga-tihan* ‘to announce, tell’, **taik-n* ‘token’, **taik-njan* ‘to show, to manifest’; (b) **taih-wō* ‘toe’:

(a) Old Norse *tjā* ‘to show’; Old High German *zeig-ōn* ‘to show’, *zih-an* ‘to accuse’, *zeihh-an* ‘sign’, *ziht* ‘accusation’; Old Franconian *teik-in* ‘sign’; Old Frisian *tig-ia* ‘to show’, *tēk-an* ‘sign’; Old English *tē-on* ‘to show’, *tāc-an* ‘to teach’, *tāc-en* ‘sign’; Middle High German *zeig-en* ‘to show’, *zih-en* ‘to accuse’, *zeich-en* ‘sign, example’; Middle Low German *tie-n* (participle *tig-en*) ‘to show’, *tēk-en* ‘sign’; Middle Dutch *tie-n* ‘to show’, *tēk-en* ~ *teik-en* ‘sign’, *tiht* ‘accusation’; Icelandic *tig-n* ~ *teik-na* ‘give a sign’, Faeroese *tek-na* ‘to show’, *tek-n* ‘sign’; Norwegian *te* ‘to show’, *teik-n* ‘sign’; Swedish *te* ‘to show’, *teck-en* ‘sign’; Danish *te* ‘to show’, *teg-n* ‘sign’; English *teach*, *tok-en*; Dutch *aan-tijg-en* ~ *op-tijg-en* ~ *be-tijg-en* ‘to show’, *tek-en* ‘sign’, dial. *teiken* ‘sign’; German *zeig-en* ‘to show’, *Zeich-en* ‘sign’, Alsatian *zaig-a* ‘to show’, *Zaich-a* ‘sign’.

(b) Old Norse *tā* 'toe'; Old High German *zēh-a* 'toe'; Old English *tāh-e* 'toe'; Middle Low German *tēwe* 'toe'; Icelandic *tá* 'toe'; Faeroese *tá* 'toe'; Norwegian *tå* 'toe'; Swedish *tå* 'toe'; Danish *tå* 'toe'; English *toe*; Dutch *teen* 'toe'; German *Zehe* 'toe'; Alsatian *Zeche* 'toe'.

Baltic: *Proto-Baltic* **teig-* 'to tell'; Old Lithuanian *tieg* 'he said'; Lithuanian *téig-ti* 'to say, tell, claim'.

Italic: (a) Oscan *deik-um* 'to say'; Umbrian *tik-amne* 'dedicace'; Latin *dic-ere* 'to say', *dic-tiō* 'discourse', *ju-dex* 'judge' (telling *ju-s*, the law); French *di-re* 'to say', *in-diqu-er* 'to show'; Occitan *dīs-er* ~ *digu-er* 'to say'; Catalan *di-r* 'to say'; Aragonese *dí* 'to say'; Spanish *dec-ir* 'to say'; Portuguese *diz-er* 'to say'; Sursilvan *di-r* 'to say'; Sutsilvan *gi-r* [džir] 'to say'; Surmiran *dei-r* 'to say'; Putér *di-r* 'to say'; Vallader *di-r* 'to say'; Friulian *dí* 'to say'; Italian *di-re* 'to say'; Romanian *zic-e* 'to say'.

(b) Latin *in-dex* 'indicative, index finger', *dig-itus* 'finger'; French *doig-t* 'finger'; Occitan *de-t* 'finger', *en-dèi-s* 'index finger'; Catalan *di-t* 'finger'; Spanish *de-do* 'finger'; Portuguese *de-do* 'finger'; Sursilvan *de-t* 'finger'; Sutsilvan *de-t* 'finger'; Surmiran *de-t* 'finger'; Putér *dau-nt* 'finger'; Vallader *dai-nt* 'finger'; Friulian *dê-t* 'finger'; Italian *di-to* 'finger'; Romanian *deg-et* 'finger'.

2. *Proto-Indo-European* **dekṃ-* 'ten' (all reflexes below also mean 'ten'):

Indic: *Proto-Indic* **dasan*; Vedic *dāça*; Prakrit *dasa* ~ *daha*; Pali *dasa*; Aśokan *daśa* ~ *dasa*; Apabhramśa *dasa* ~ *daha*; European Romany *deš*; Armenia Romany *las*; Palestine Romany *das*; Gondwani *dhamak*; Dameli *daš*; Domaki *dai*; Tirahi *dā*; Poguli *dāh*; Rambani *das*; Kohistani *daš*; Pashai *dāya*; Shumashti *dās*; Ningalami *das*; Wotapuri *daš*; Gawarbatī *dōš*; Kalasha *daš*; Khowar *još*; Bashkarik *daš*; Torwali *daš*; Kandia *daš*; Maiyā *daš*; Savi *daš*; Phalula *dāš*; Shina *dāi*; Kashmiri *dah*; Rambani *das*; Poguli *dāh*; Dodi *dās*; Sindhi *dāha*; Khatri *dō*; Kacchi *dau*; Lahnda *dāh*; Khetrani *dā*; Awankari *dā*; Punjabi *das*; Siraiki *dah*; Western Pahari *daš*; Kotgarhi *dōš*; Garhwali *das*; Kumauni *das*; Nepali *das*; Assamese *dah*; Mayang *dos*; Bengali *das*; Oriya *dasa*; Bihari *das*; Maithili *das* ~ *dah*; Magahi *das*; Bhojpuri *das*; Awadi *das*; Lakhimpuri *das*; Hindi *das*; Bhili *dōh* ~ *dah*; Dogri *das*; Chattisgarhi *das*; Khandeshi *das*; Braj *das*; Bundeli *das*; Urdu *das*; Rajasthani *das*; Malvi *das*; Magaji *das*; Marwari *das*; Gujarati *das*; Marathi *das* ~ *dahā*; Konkani *dhā*; Sinhalese *dasa-ya* ~ *daha-ya*; Maldivian *diha*. (Turner 1962–1966: etym. 6227; Rosenfelder no date)

Nuristani: Kalasha-ala (Waigali) *dōš*; Wasi-weri *lez*; Kati *duc*; Kamviri *d'uç*; Ashkun *dus*. (Turner 1962–1966: etym. 6227)

Iranian: Avestan *dasa*; Pahlavi *dah*; Khotanese *dasau*; Khwarezmian *dhs*; Turfanian *dh*; Iron Ossetic *dæs*; Digor Ossetic *dæs*; Yaghnobi *das*; Pashto *las*; Wakhi *das*; Munji *dah*; Ishkashmi *da*; Sanglechi *das*; Zebaki *dos*; Shughni *dis*; Yidgha *los*; Rushani *des*; Yazgulami *ḍus*; Sarikoli *des*; Parachi *dōs*; Ormuri *das*; Nayini *de*; Natanzi *dē*; Khunsari *dē*; Gazi *dē*; Sivandi *da*; Vafsi *dah*; Semnani *das*; Gilaki *da*; Mazanderani *da*; Talysh *dā*; Harzani *doh*; Zaza *des*; Gorani *da*; Balochi *dah*; Southern Kurdish *da*; Northern Kurdish *da*; Persian *dah*; Tajik *dah*; Tati *dæh*; Chali *dā*; Farsi *daśa*; Lari *da*; Luri *dah*; Kumzari *da'hata*.

Tocharian: Tocharian A *säk*; Tocharian B *śak*.

Armenian: Classical Armenian *t'asn*; Western Armenian *tas*.

Hellenic: Classical Attic Greek *dēka*; Aeolic *dēko*; Modern Greek *ḗka*; Tsakonian *ḗka*; Cypriot *ḗga*; Pontic *ḗka* ~ *rēka*.

- Slavic: Proto-Slavic **des-ęt*ʹ; Old Church Slavonic *des-ętŭ*; Russian *dés-jat*ʹ; Belorussian *dzés-jac*ʹ; Ukrainian *dés-jat*ʹ; Polish *dzies-ięć*; Kashubian *dzes-ińc*; Polabian *dis-qt*; Czech *des-et*; Slovak *des-at*ʹ; Eastern Slovak *dzeš-ec*; Upper Sorbian *džes-ać*; Lower Sorbian *žas-eś*; Bulgarian *dés-et*; Serbo-Croatian *dēs-ët*; Slovene *des-et*; Macedonian *des-et*.
- Baltic: Proto-Baltic **dēčim-t-*; Old Prussian *dessim-pts*; Lithuanian *dēšim-tis*; Latvian *desm-its*.
- Germanic: Proto-Germanic **tíxun*; Gothic *taíhun*; Old Norse *tíu*; Old Icelandic *tíō*; Old Swedish *tíō*; Old Danish *tí*; Old High German *zehan*; Old Saxon *tehan*; Old Frisian *tian*; Old Low Franconian *tēn*; Old English *tēne*; Middle Low German *tein*; Middle Dutch *thien*; Middle High German *zehen*; Icelandic *tío*; Norwegian *tio* ~ *tie*; Swedish *tio*; Dalecarlian *tiu*; Faeroese *tíggju*; West Frisian *tsien*; Saterland Frisian *tjoon*; Fohr North Frisian *tjiin*; Sylt North Frisian *tiin*; Helgoland North Frisian *tain*; Dutch *tien*; Low Saxon *tain*; Westphalian Saxon *tein*; Crimean Gothic *thiine*; English *ten*; German *zehn*; Bavarian *zene*; Swabian *zaen*; Cimbrian *zègan*; Rhine Franconian *zeen*; Luxemburgish *zèng*; Swiss German *zäh*.
- Italic: Latin *decem*; Old French *dis*; French *dix*; Walloon *dijh*; Jèrriais *dgix*; Picard *dich*; Poitevin *dis*; Occitan *dètz*; North Occitan *dié*; Franco-Provençal *dyî*; Aragonese *deu*, *dech-igüeit* 'eighteen'; Catalan *deu*; Spanish *diez*; Ladino *dies*; Asturian *diez*; Galician *dez*; Portuguese *dez*; Sursilvan *diesch*; Sutsilvan *diesch*; Vallader *desch*; Friulian *dis*; Ladin *diesc*; Piedmontese *dés*; Milanese *dés*; Genovese *dexe*; Venetian *diese*; Corsican *dece*; Umbrian *dèsce*; Neapolitan *riècë*; Sicilian *dècis*; Italian *dieci*; Sardinian *deghe*; Romanian *zece*; Arumanian *date*; Meglenian *zeti*.
- Celtic: Gaulish *decam*; Old Irish *deich*; Irish *deich*; Scottish Gaelic *deich*; Manx *jeih*; Welsh *deg*; Breton *dek*; Vannetais *dek*; Cornish *dek*.
- Albanian: Standard Albanese *dhjetë*; Gheg *ðet*; Tosk *zjëtë*.

